(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 September 2003 (25.09.2003)

PCT

(10) International Publication Number
WO 03/078662 A1

(51) International Patent Classification[7]: C12Q 1/68, G01N 33/53

(21) International Application Number: PCT/US03/07713

(22) International Filing Date: 12 March 2003 (12.03.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/364,890 13 March 2002 (13.03.2002) US
60/412,049 18 September 2002 (18.09.2002) US

(71) Applicant *(for all designated States except US)*: **GENOMIC HEALTH** [US/US]; 301 Penobscot Drive, Redwood City, CA 94063 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BAKER, Joffre, B.** [US/US]; P.O. Box 371212, Montara,CA 94937 (US). **CRONIN, Maureen, T.** [US/US]; 771 Anderson Drive, Los Altos, CA 94024 (US). **KIEFER, Michael, C.** [US/US]; 401 Wright Court, Clayton, CA 94517 (US). **SHAK, Steve** [US/US]; 1133 Cambridge Road, Burlingame, CA 94010 (US). **WALKER, Michael, G.** [US/US]; 1475 Flamingo Way, Sunnyvale, CA 94087 (US).

(74) Agent: **DREGER, Ginger, R.**; Heller Ehrman White & McAuliffe, 275 Middlefield Road, Menlo Park, CA 94025-3506 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 03/078662 A1

(54) Title: GENE EXPRESSION PROFILING IN BIOPSIED TUMOR TISSUES

(57) Abstract: The invention concerns sensitive methods to measure mRNA levels in biopsied tumor tissues, including archived paraffin-embedded biopsy material. Th invention also concerns breast cancer gene sets important in the diagnosis and treatment of breast cancer, and methods for assigning the most optimal treatment options to breast cancer patient based upon knowledge derived from gene expression studies.

GENE EXPRESSION PROFILING IN BIOPSIED TUMOR TISSUES

Background of the Invention

Cross-Reference

This application claims the benefit under 35 U.S.C. 119(h) of provisional applications serial nos. 60/412,049, filed September 18, 2002 and 60/364,890, filed March 13, 2002, the entire disclosures which are herby incorporated by reference.

Field of the Invention

The present invention relates to gene expression profiling in biopsied tumor tissues. In particular, the present invention concerns sensitive methods to measure mRNA levels in biopsied tumor tissues, including archived paraffin-embedded biopsy material. In addition, the invention provides a set of genes the expression of which is important in the diagnosis and treatment of breast cancer.

Oncologists have a number of treatment options available to them, including different combinations of chemotherapeutic drugs that are characterized as "standard of care," and a number of drugs that do not carry a label claim for a particular cancer, but for which there is evidence of efficacy in that cancer. Best likelihood of good treatment outcome requires that patients be assigned to optimal available cancer treatment, and that this assignment be made as quickly as possible following diagnosis.

Currently, diagnostic tests used in clinical practice are single analyte, and therefore do not capture the potential value of knowing relationships between dozens of different markers. Moreover, diagnostic tests are frequently not quantitative, relying on immunohistochemistry. This method often yields different results in different laboratories, in part because the reagents are not standardized, and in part because the interpretations are subjective and cannot be easily quantified. RNA-based tests have not often been used because of the problem of RNA degradation over time and the fact that it is difficult to obtain fresh tissue samples from patients for analysis. Fixed paraffin-embedded tissue is more readily available and methods have been established to detect RNA in fixed tissue. However, these methods typically do not allow for the study of large numbers of genes (DNA or RNA) from small amounts of material. Thus, traditionally fixed tissue has been rarely used other than for immunohistochemistry detection of proteins.

Recently, several groups have published studies concerning the classification of various cancer types by microarray gene expression analysis (see, e.g. Golub *et al.*, *Science* 286:531-537

(1999); Bhattacharjae *et al., Proc. Natl. Acad. Sci. USA* 98:13790-13795 (2001); Chen-Hsiang *et al., Bioinformatics* 17 (Suppl. 1):S316-S322 (2001); Ramaswamy *et al., Proc. Natl. Acad. Sci. USA* 98:15149-15154 (2001)). Certain classifications of human breast cancers based on gene expression patterns have also been reported (Martin *et al., Cancer Res.* 60:2232-2238 (2000); West *et al., Proc. Natl. Acad. Sci. USA* 98:11462-11467 (2001); Sorlie *et al., Proc. Natl. Acad. Sci. USA* 98:10869-10874 (2001); Yan *et al., Cancer Res.* 61:8375-8380 (2001)). However, these studies mostly focus on improving and refining the already established classification of various types of cancer, including breast cancer, and generally do not provide new insights into the relationships of the differentially expressed genes, and do not link the findings to treatment strategies in order to improve the clinical outcome of cancer therapy.

Although modern molecular biology and biochemistry have revealed more than 100 genes whose activities influence the behavior of tumor cells, state of their differentiation, and their sensitivity or resistance to certain therapeutic drugs, with a few exceptions, the status of these genes has not been exploited for the purpose of routinely making clinical decisions about drug treatments. One notable exception is the use of estrogen receptor (ER) protein expression in breast carcinomas to select patients to treatment with anti-estrogen drugs, such as tamoxifen. Another exceptional example is the use of ErbB2 (Her2) protein expression in breast carcinomas to select patients with the Her2 antagonist drug Herceptin® (Genentech, Inc., South San Francisco, CA).

Despite recent advances, the challenge of cancer treatment remains to target specific treatment regimens to pathogenically distinct tumor types, and ultimately personalize tumor treatment in order to maximize outcome. Hence, a need exists for tests that simultaneously provide predictive information about patient responses to the variety of treatment options. This is particularly true for breast cancer, the biology of which is poorly understood. It is clear that the classification of breast cancer into a few subgroups, such as $ErbB2^+$ subgroup, and subgroups characterized by low to absent gene expression of the estrogen receptor (ER) and a few additional transcriptional factors (Perou *et al., Nature* 406:747-752 (2000)) does not reflect the cellular and molecular heterogeneity of breast cancer, and does not allow the design of treatment strategies maximizing patient response.

Summary of the Invention

The present invention provides (1) sensitive methods to measure mRNA levels in biopsied tumor tissue, (2) a set of approximately 190 genes, the expression of which is important in the diagnosis of breast cancer, and (3) the significance of abnormally low or high expression

for the genes identified and included in the gene set, through activation or disruption of biochemical regulatory pathways that influence patient response to particular drugs used or potentially useful in the treatment of breast cancer. These results permit assessment of genomic evidence of the efficacy of more than a dozen relevant drugs.

The present invention accommodates the use of archived paraffin-embedded biopsy material for assay of all markers in the set, and therefore is compatible with the most widely available type of biopsy material. The invention presents an efficient method for extraction of RNA from wax-embedded, fixed tissues, which reduces cost of mass production process for acquisition of this information without sacrificing quality of the analysis. In addition, the invention describes a novel highly effective method for amplifying mRNA copy number, which permits increased assay sensitivity and the ability to monitor expression of large numbers of different genes given the limited amounts of biopsy material. The invention also captures the predictive significance of relationships between expressions of certain markers in the breast cancer marker set. Finally, for each member of the gene set, the invention specifies the oligonucleotide sequences to be used in the test.

In one aspect, the invention concerns a method for predicting clinical outcome for a patient diagnosed with cancer, comprising

determining the expression level of one or more genes, or their expression products, selected from the group consisting of p53BP2, cathepsin B, cathepsin L, Ki67/MiB1, and thymidine kinase in a cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference cancer tissue set,

wherein a poor outcome is predicted if:

(a) the expression level of p53BP2 is in the lower 10^{th} percentile; or

(b) the expression level of either cathepsin B or cathepsin L is in the upper 10^{th} percentile; or

(c) the expression level of any either Ki67/MiB1 or thymidine kinase is in the upper 10^{th} percentile.

Poor clinical outcome can be measured, for example, in terms of shortened survival or increased risk of cancer recurrence, e.g. following surgical removal of the cancer.

In another embodiment, the inventor concerns a method of predicting the likelihood of the recurrence of cancer, following treatment, in a cancer patient, comprising determining the expression level of p27, or its expression product, in a cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference

cancer tissue set, wherein an expression level in the upper 10th percentile indicates decreased risk of recurrence following treatment.

In another aspect, the invention concerns a method for classifying cancer comprising, determining the expression level of two or more genes selected from the group consisting of Bcl2, hepatocyte nuclear factor 3, ER, ErbB2, and Grb7, or their expression products, in a cancer tissue, normalized against a control gene or genes, and compared to the amount found in a reference cancer tissue set, wherein (i) tumors expressing at least one of Bcl2, hepatocyte nuclear factor 3, and ER, or their expression products, above the mean expression level in the reference tissue set are classified as having a good prognosis for disease free and overall patient survival following treatment; and (ii) tumors expressing elevated levels of ErbB2 and Grb7, or their expression products, at levels ten-fold or more above the mean expression level in the reference tissue set are classified as having poor prognosis of disease free and overall patient survival following treatment.

All types of cancer are included, such as, for example, breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer, pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, and brain cancer. The foregoing methods are particularly suitable for prognosis/classification of breast cancer.

In all previous aspects, in a specific embodiment, the expression level is determined using RNA obtained from a formalin-fixed, paraffin-embedded tissue sample. While all techniques of gene expression profiling, as well as proteomics techniques, are suitable for use in performing the foregoing aspects of the invention, the gene expression levels are often determined by reverse transcription polymerase chain reaction (RT-PCR).

If the source of the tissue is a formalin-fixed, paraffin embedded tissue sample, the RNA is often fragmented.

The expression data can be further subjected to multivariate analysis, for example using the Cox Proportional Hazards model.

In a further aspect, the invention concerns a method for the preparation of nucleic acid from a fixed, wax-embedded tissue specimen, comprising:

(a) incubating a section of the fixed, wax-embedded tissue specimen at a temperature of about 56 °C to 70 °C in a lysis buffer, in the presence of a protease, without prior dewaxing, to form a lysis solution;

(b) cooling the lysis solution to a temperature where the wax solidifies; and

(c) isolating the nucleic acid from the lysis solution.

The lysis buffer may comprise urea, such as 4M urea.

In a particular embodiment, incubation in step (a) of the foregoing method is performed at about 65°C.

In another particular embodiment, the protease used in the foregoing method is proteinase K.

In another embodiment, the cooling in step (b) is performed at room temperature.

In a further embodiment, the nucleic acid is isolated after protein removal with 2.5 M NH_4OAc.

The nucleic acid can, for example, be total nucleic acid present in the fixed, wax-embedded tissue specimen.

In yet another embodiment, the total nucleic acid is isolated by precipitation from the lysis solution, following protein removal, with 2.5 M NH_4OAc. The precipitation may, for example, be performed with isopropanol.

The method described above may further comprise the step of removing DNA from the total nucleic acid, for example by DNAse treatment.

The tissue specimen may, for example, be obtained from a tumor, and the RNA may be obtained from a microdissected portion of the tissue specimen enriched for tumor cells.

All types of tumor are included, such as, without limitation, breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer, pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, and brain cancer, in particular breast cancer..

The method described above may further comprise the step of subjecting the RNA to gene expression profiling. Thus, the gene expression profile may be completed for a set of genes comprising at least two of the genes listed in Table 1.

Although all methods of gene expression profiling are contemplated, in a particular embodiment, gene expression profiling is performed by RT-PCR which may be preceded by an amplification step.

In another aspect, the invention concerns a method for preparing fragmented RNA for gene expression analysis, comprising the steps of:

(a) mixing the RNA with at least one gene-specific, single-stranded DNA scaffold under conditions such that fragments of the RNA complementary to the DNA scaffold hybridize with the DNA scaffold;

(b) extending the hybridized RNA fragments with a DNA polymerase to form a DNA-DNA duplex; and

(c) removing the DNA scaffold from the duplex.

In a specific embodiment, in step (b) of this method, the RNA may be mixed with a mixture of single-stranded DNA templates specific for each gene of interest.

The method can further comprise the step of heat-denaturing and reannealing the duplexed DNA to the DNA scaffold, with or without additional overlapping scaffolds, and further extending the duplexed sense strand with DNA polymerase prior to removal of the scaffold in step (c).

The DNA templates may be, but do not need to be, fully complementary to the gene of interest.

In a particular embodiment, at least one of the DNA templates is complementary to a specific segment of the gene of interest.

In another embodiment, the DNA templates include sequences complementary to polymorphic variants of the same gene.

The DNA template may include one or more dUTP or rNTP sites. In this case. iin step (c) the DNA template may be removed by fragmenting the DNA template present in the DNA-DNA duplex formed in step (b) at the dUTP or rNTP sites.

In an important embodiment, the RNA is extracted from fixed, wax-embedded tissue specimens, and purified sufficiently to act as a substrate in an enzyme assay. The RNA purification may, but does not need to, include an oligo-dT based step.

In a further aspect, the invention concerns a method for amplifying RNA fragments in a sample comprising fragmented RNA representing at least one gene of interest, comprising the steps of:

(a) contacting the sample with a pool of single-stranded DNA scaffolds comprising an RNA polymerase promoter at the 5' end under conditions such that the RNA fragments complementary to the DNA scaffolds hybridize with the DNA scaffolds;

(b) extending the hybridized RNA fragments with a DNA polymerase along the DNA scaffolds to form DNA-DNA duplexes;

(c) amplifying the gene or genes of interest by *in vitro* transcription; and

(d) removing the DNA scaffolds from the duplexes.

An exemplary promoter is the T7 RNA polymerase promoter, while an exemplary DNA polymerase is DNA polymerase I.

In step (d) the DNA scaffolds may be removed, for example, by treatment with DNase I.

In a further embodiment, the pool of single-stranded DNA scaffolds comprises partial or complete gene sequences of interest, such as a library of cDNA clones.

In a specific embodiment, the sample represents a whole genome or a fraction thereof. In a preferred embodiment, the genome is the human genome.

In another aspect, the invention concerns a method of preparing a personalized genomics profile for a patient, comprising the steps of:

(a) subjecting RNA extracted from a tissue obtained from the patient to gene expression analysis;

(b) determining the expression level in such tissue of at least two genes selected from the gene set listed in Table 1, wherein the expression level is normalized against a control gene or genes, and is compared to the amount found in a cancer tissue reference set;

(c) and creating a report summarizing the data obtained by the gene expression analysis.

The tissue obtained from the patient may, but does not have to, comprise cancer cells. Just as before, the cancer can, for example, be breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer, pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, or brain cancer, breast cancer being particularly preferred.

In a particular embodiment, the RNA is obtained from a microdissected portion of breast cancer tissue enriched for cancer cells. The control gene set may, for example, comprise S-actin, and ribosomal protein LPO.

The report prepared for the use of the patient or the patient's physician, may include the identification of at least one drug potentially beneficial in the treatment of the patient.

Step (b) of the foregoing method may comprise the step of determining the expression level of a gene specifically influencing cellular sensitivity to a drug, where the gene can, for example, be selected from the group consisting of aldehyde dehydrogenase 1A1, aldehyde dehydrogenase 1A3, amphiregulin, ARG, BRK, BCRP, CD9, CD31, CD82/KAI-1, COX2, c-abl, c-kit, c-kit L, CYP1B1, CYP2C9, DHFR, dihydropyrimidine dehydrogenase, EGF, epiregulin, ER-alpha, ErbB-1, ErbB-2, ErbB-3, ErbB-4, ER-beta, farnesyl pyrophosphate synthetase, gamma-GCS (glutamyl cysteine synthetase), GATA3, geranyl geranyl pyrophosphate synthetase, Grb7, GST-alpha, GST-pi, HB-EGF, hsp 27, human chorionic gonadotropin/CGA, IGF-1, IGF-2, IGF1R, KDR, LIV1, Lung Resistance Protein/MVP, Lot1, MDR-1, microsomal epoxide hydrolase, MMP9, MRP1, MRP2, MRP3, MRP4, PAI1, PDGF-A, PDGF-B, PDGF-C, PDGF-D, PGDFR-alpha, PDGFR-beta, PLAGa (pleiomorphic adenoma 1), PREP prolyl endopeptidase, progesterone receptor, pS2/trefoil factor 1, PTEN, PTB1b, RAR-alpha, RAR-beta2, Reduced

Folate Carrier, SXR, TGF-alpha, thymidine phosphorylase, thymidine synthase, topoisomerase II-alpha, topoisomerase II-beta, VEGF, XIST, and YB-1.

In another embodiment, step (b) of the foregoing process includes determining the expression level of multidrug resistance factors, such as, for example, gamma-glutamyl-cysteine synthetase (GCS), GST-α, GST-π, MDR-1, MRP1-4, breast cancer resistance protein (BCRP), lung cancer resistance protein (MVP), SXR, or YB-1.

In another embodiment, step (b) of the foregoing process comprises determination of the expression level of eukaryotic translation initiation factor 4E (EIF4E).

In yet another embodiment, step (b) of the foregoing process comprises determination of the expression level of a DNA repair enzyme.

In a further embodiment, step (b) of the foregoing process comprises determination of the expression level of a cell cycle regulator, such as, for example, c-MYC, c-Src, Cyclin D1, Ha-Ras, mdm2. p14ARF, p21WAF1/CI, p16INK4a/p14, p23, p27, p53, PI3K, PKC-epsilon, or PKC-delta.

In a still further embodiment, step (b) of the foregoing process comprises determination of the expression level of a tumor suppressor or a related protein, such as, for example, APC or E-cadherin.

In another embodiment, step (b) of the foregoing method comprises determination of the expression level of a gene regulating apoptosis, such as, for example, p53, BC12, Bcl-x1, Bak, Bax, and related factors, NFκ-B, CIAP1, CIAP2, survivin, and related factors, p53BP1/ASPP1, or p53BP2/ASPP2.

In yet another embodiment, step (b) of the foregoing process comprises determination of the expression level of a factor that controls cell invasion or angiogenesis, such as, for example, uPA, PAI1, cathepsin B, C, and L, scatter factor (HGF), c-met, KDR, VEGF, or CD31.

In a different embodiment, step (b) of the foregoing method comprises determination of the expression level of a marker for immune or inflammatory cells or processes, such as, for example, Ig light chain λ, CD18, CD3, CD68. Fas(CD95), or Fas Ligand.

In a further embodiment, step (b) of the foregoing process comprises determination of the expression level of a cell proliferation marker, such as, for example, Ki67/MiB1, PCNA, Pin1, or thymidine kinase.

In a still further embodiment, step (b) of the foregoing process comprises determination of the expression level of a growth factor or growth factor receptor., such as, for example, IGF1, IGF2, IGFBP3, IGF1R, FGF2, CSF-1, CSF-1R/fms, SCF-1, IL6 or IL8.

In another embodiment, step (b) of the foregoing process comprises determination of the expression level of a gene marker that defines a subclass of breast cancer, where the gene marker can, for example, be GRO1 oncogene alpha, Grb7, cytokeratins 5 and 17, retinol binding protein 4, hepatocyte nuclear factor 3, integrin subunit alpha 7, or lipoprotein lipase.

In a still further aspect, the invention concerns a method for predicting the response of a patient diagnosed with breast cancer to 5-fluorouracil (5-FU) or an analog thereof, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis;

(b) determining the expression level in the tissue of thymidylate synthase mRNA, wherein the expression level is normalized against a control gene or genes, and is compared to the amount found in a reference breast cancer tissue set; and

(c) predicting patient response based on the normalized thymidylate synthase mRNA level.

Step (d) of the foregoing method can further comprise determining the expression level of dihydropyrimidine phosphorylase.

In another embodiment, step (b) of the method can further comprise determining the expression level of thymidine phosphorylase.

In yet another embodiment, a positive response to 5-FU or an analog thereof is predicted if: (i) normalized thymidylate synthase mRNA level determined in step (b) is at or below the 15th percentile; or (ii) the sum of normalized expression levels of thymidylate synthase and dihydropyrimidine phosphorylase determined in step (b) is at or below the 25th percentile; or (iii) the sum of normalized expression levels of thymidylate synthase, dihydropyrimidine phosphorylase, plus thymidine phosphorylase determined in step (b) is at or below the 20th percentile..

In a further embodiment, in step (b) of the foregoing method the expression level of c-myc and wild-type p53 is determined. In this case, a positive response to 5-FU or an analog thereof is predicted, if the normalized expression level of c-myc relative to the normalized expression level of wild-type p53 is in the upper 15th percentile.

In a still further embodiment, in step (b) of the foregoing method, expression level of NFκB and cIAP2 is determined. In this particular embodiment, resistance to 5-FU or an analog thereof is typically predicted if the normalized expression level of NFκB and cIAP2 is at or above the 10th percentile.

In another aspect, the invention concerns a method for predicting the response of a patient diagnosed with breast cancer to methotrexate or an analog thereof, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting decreased patient sensitivity to methotrexate or analog if (i) DHFR levels are more than tenfold higher than the average expression level of DHFR in the control gene set, or (ii) the normalized expression levels of members of the reduced folate carrier (RFC) family are below the 10^{th} percentile.

In yet another aspect, the invention concerns a method for predicting the response of a patient diagnosed with breast cancer to an anthracycline or an analog thereof, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting patient resistance or decreased sensitivity to the anthracycline or analog if (i) the normalized expression level of topoisomerase IIα is below the 10^{th} percentile, or (ii) the normalized expression level of topoisomerase IIβ is below the 10^{th} percentile, or (iii) the combined normalized topoisomerase IIα or IIβ expression levels are below the 10^{th} percentile.

In a different aspect, the invention concerns a method for predicting the response of a patient diagnosed with breast cancer to a docetaxol, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting reduced sensitivity to docetaxol if the normalized expression level of CYP1B1 is in the upper 10^{th} percentile.

The invention further concerns a method for predicting the response of a patient diagnosed with breast cancer to cyclophosphamide or an analog thereof, comprising

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting reduced sensitivity to the cyclophosphamide or analog if the sum of the expression levels of aldehyde dehydrogenase 1A1 and 1A3 is more than tenfold higher than the average of their combined expression levels in the reference tissue set.

In a further aspect, the invention concerns a method for predicting the response of a patient diagnosed with breast cancer to anti-estrogen therapy, comprising

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set that contains both specimens negative for and positive for estrogen receptor-α (ERα) and progesterone receptor-α (PRα); and

(b) predicting patient response based upon the normalized expression levels of ERα or PRα, and at least one of microsomal epoxide hydrolase, pS2/trefoil factor 1, GATA3 and human chorionic gonadotropin.

In a specific embodiment, lack of response or decreased responsiveness is predicted if (i) the normalized expression level of microsomal epoxide hydrolase is in the upper 10th percentile; or (ii) the normalized expression level of pS2/trefoil factor 1, or GATA3 or human chorionic gonaostropin is at or below the corresponding average expression level in said breast cancer tissue set, regardless of the expression level of ERα or PRα in the breast cancer tissue obtained from the patient.

In another aspect, the invention concerns a method for predicting the response of a patient diagnosed with breast cancer to a taxane, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting reduced sensitivity to taxane if (i) no or minimal XIST expression is detected; or (ii) the normalized expression level of GST-π or propyl endopeptidase (PREP) is in the upper 10th percentile; or (iii) the normalized expression level of PLAG1 is in the upper 10th percentile.

The invention also concerns a method for predicting the response of a patient diagnosed with breast cancer to cisplatin or an analog thereof, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting resistance or reduced sensitivity if the normalized expression level of ERCC1 is in the upper 10th percentile.

The invention further concerns a method for predicting the response of a patient diagnosed with breast cancer to an ErbB2 or EGFR antagonist, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting patient response based on the normalized expression levels of at least one of Grb7, IGF1R, IGF1 and IGF2.

In particular embodiment, a positive response is predicted if the normalized expression level of Grb7 is in the upper 10^{th} percentile, and the expression of IGF1R, IGF1 and IGF2 is not elevated above the 90^{th} percentile.

In a further particular embodiment, a decreased responsiveness is predicted if the expression level of at least one of IGF1R, IGF1 and IGF2 is elevated.

In another aspect, the invention concerns a method for predicting the response of a patient diagnosed with breast cancer to a bis-phosphonate drug, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting a positive response if the breast cancer tissue obtained from the patient expresses mutant Ha-Ras and additionally expresses farnesyl pyrophosphate synthetase or geranyl pyrophosphone synthetase at a normalized expression level at or above the 90^{th} percentile.

In yet another aspect, the invention concerns a method for predicting the response of a patient diagnosed with breast cancer to treatment with a cyclooxygenase 2 inhibitor, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting a positive response if the normalized expression level of COX2 in the breast cancer tissue obtained from the patient is at or above the 90^{th} percentile.

The invention further concerns a method for predicting the response of a patient diagnosed with breast cancer to an EGF receptor (EGFR) antagonist , comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting a positive response to an EGFR antagonist, if (i) the normalized expression level of EGFR is at or above the 10^{th} percentile, and (ii) the normalized expression

level of at least one of epiregulin, TGF-α, amphiregulin, ErbB3, BRK, CD9, MMP9, CD82, and Lot1 is above the 90th percentile.

In another aspect, the invention concerns a method for monitoring the response of a patient diagnosed with breast cancer to treatment with an EGFR antagonist, comprising monitoring the expression level of a gene selected from the group consisting of epiregulin, TGF-α, amphiregulin, ErbB3, BRK, CD9, MMP9, CD82, and Lot1 in the patient during treatment, wherein reduction in the expression level is indicative of positive response to such treatment.

In yet another aspect, the invention concerns a method for predicting the response of a patient diagnosed with breast cancer to a drug targeting a tyrosine kinase selected from the group consisting of abl, c-kit, PDGFR-α, PDGFR-β and ARG , comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set;

(b) determining the normalized expression level of a tyrosine kinase selected from the group consisting of abl, c-kit, PDGFR-α, PDGFR-β and ARG, and the cognate ligand of the tyrosine kinase, and if the normalized expression level of the tyrosine kinase is in the upper 10th percentile,

(c) determining whether the sequence of the tyrosine kinase contains any mutation,

wherein a positive response is predicted if (i) the normalized expression level of the tyrosine kinase is in the upper 10th percentile, (ii) the sequence of the tyrosine kinase contains an activating mutation, or (iii) the normalized expression level of the tyrosine kinase is normal and the expression level of the ligand is in the upper 10th percentile.

Another aspect of the invention is a method for predicting the response of a patient diagnosed with breast cancer to treatment with an anti-angiogenic drug, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) predicting a positive response if (i) the normalized expression level of VEGF is in the upper 10th percentile and (ii) the normalized expression level of KDR or CD31 is in the upper 20th percentile.

A further aspect of the invention is a method for predicting the likelihood that a patient diagnosed with breast cancer develops resistance to a drug interacting with the MRP-1 gene coding for the multidrug resistance protein P-glycoprotein, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis to determine the expression level of PTP1b, wherein the expression level is normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) concluding that the patient is likely to develop resistance to said drug if the normalized expression level of the MRP-1 gene is above the 90th percentile.

The invention further relates to a method for predicting the likelihood that a patient diagnosed with breast cancer develops resistance to a chemotherapeutic drug or toxin used in cancer treatment, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) determining the normalized expression levels of at least one of the following genes: MDR1, SGTα, GSTπ, SXR, BCRP YB-1, and LRP/MVP, wherein the finding of a normalized expression level in the upper 4th percentile is an indication that the patient is likely to develop resistance to the drug.

Also included herein is a method for measuring the translational efficiency of VEGF mRNA in a breast cancer tissue sample, comprising determining the expression levels of the VEGF and EIF4E mRNA in the sample, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein a higher normalized EIF4E expression level for the same VEGF expression level is indicative of relatively higher translational efficiency for VEGF.

In another aspect, the invention provides a method for predicting the response of a patient diagnosed with breast cancer to a VEGF antagonist, comprising determining the expression level of VEGF and EIF4E mRNA normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein a VEGF expression level above the 90th percentile and an EIF4E expression level above the 50th percentile is a predictor of good patient response.

The invention further provides a method for predicting the likelihood of the recurrence of breast cancer in a patient diagnosed with breast cancer, comprising determining the ratio of p53:p21 mRNA expression or p53:mdm2 mRNA expression in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein an above normal ratio is indicative of a higher risk

of recurrence. Typically, a higher risk of recurrence is indicated if the ratio is in the upper 10th percentile.

In yet another aspect, the invention concerns a method for predicting the likelihood of the recurrence of breast cancer in a breast cancer patient following surgery, comprising determining the expression level of cyclin D1 in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein an expression level in the upper 10th percentile indicates increased risk of recurrence following surgery. In a particular embodiment of this method, the patient is subjected to adjuvant chemotherapy, if the expression level is in the upper 10th percentile.

Another aspect of the invention is a method for predicting the likelihood of the recurrence of breast cancer in a breast cancer patient following surgery, comprising determining the expression level of APC or E-cadherin in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein an expression level in the upper 5th percentile indicates high risk of recurrence following surgery, and heightened risk of shortened survival.

A further aspect of the invention is a method for predicting the response of a patient diagnosed with breast cancer to treatment with a proapoptotic drug comprising determining the expression levels of BCl2 and c-MYC in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein (i) a BCl2 expression level in the upper 10th percentile in the absence of elevated expression of c-MYC indicates good response, and (ii) a good response is not indicated if the expression level c-MYC is elevated, regardless of the expression level of BC12.

A still further aspect of the invention is a method for predicting treatment outcome for a patient diagnosed with breast cancer, comprising the steps of:

(a) subjecting RNA extracted from a breast cancer tissue obtained from the patient to gene expression analysis, wherein gene expression levels are normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set; and

(b) determining the normalized expression levels of NFκB and at least one gene selected from the group consisting of cIAP1, cIAP2, XIAP, and Survivin,

wherein a poor prognosis is indicated if the expression levels for NFκB and at least one of the genes selected from the group consisting of cIAP1, cIAP2, XIAP, and Survivin is in the upper 5th percentile.

The invention further concerns a method for predicting treatment outcome for a patient diagnosed with breast cancer, comprising determining the expression levels of p53BP1 and p53BP2 in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein a poor outcome is predicted if the expression level of either p53BP1 or p53BP2 is in the lower 10^{th} percentile.

The invention additionally concerns a method for predicting treatment outcome for a patient diagnosed with breast cancer, comprising determining the expression levels of uPA and PAI1 in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein (i) a poor outcome is predicted if the expression levels of uPA and PAI1 are in the upper 20^{th} percentile, and (ii) a decreased risk of recurrence is predicted if the expression levels of uPA and PAI1 are not elevated above the mean observed in the breast cancer reference set. In a particular embodiment, poor outcome is measured in terms of shortened survival or increased risk of cancer recurrence following surgery. In another particular embodiment, uPA and PAI1 are expressed at normal levels, and the patient is subjected to adjuvant chemotherapy following surgery.

Another aspect of the invention is a method for predicting treatment outcome in a patient diagnosed with breast cancer, comprising determining the expression levels of cathepsin B and cathepsin L in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein a poor outcome is predicted if the expression level of either cathepsin B or cathepsin L is in the upper 10^{th} percentile. Just as before, poor treatment outcome may be measured, for example, in terms of shortened survival or increased risk of cancer recurrence.

A further aspect of the invention is a method for devising the treatment of a patient diagnosed with breast cancer, comprising the steps of

(a) determining the expression levels of scatter factor and c-met in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, and

(b) suggesting prompt aggressive chemotherapeutic treatment if the expression levels of scatter factor and c-met or the combination of both, are above the 90^{th} percentile.

A still further aspect of the invention is a method for predicting treatment outcome for a patient diagnosed with breast cancer, comprising determining the expression levels of VEGF, CD31, and KDR in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein

a poor treatment outcome is predicted if the expression level of any of VEGF, CD31, and KDR is in the upper 10th percentile.

Yet another aspect of the invention is a method for predicting treatment outcome for a patient diagnosed with breast cancer, comprising determining the expression levels of Ki67/MiB1, PCNA, Pin1, and thymidine kinase in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein a poor treatment outcome is predicted if the expression level of any of Ki67/MiB1, PCNA, Pin1, and thymidine kinase is in the upper 10th percentile.

The invention further concerns a method for predicting treatment outcome for a patient diagnosed with breast cancer, comprising determining the expression level of soluble and full length CD95 in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein the presence of soluble CD95 correlates with poor patient survival.

The invention also concerns a method for predicting treatment outcome for a patient diagnosed with breast cancer, comprising determining the expression levels of IGF1, IGF1R and IGFBP3 in a breast cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein a poor treatment outcome is predicted if the sum of the expression levels of IGF1, IGF1R and IGFBP3 is in the upper 10th percentile.

The invention additionally concerns a method for classifying breast cancer comprising, determining the expression level of two or more genes selected from the group consisting of Bcl12, hepatocyte nuclear factor 3, LIV1, ER, lipoprotein lipase, retinol binding protein 4, integrin α7, cytokeratin 5, cytokeratin 17, GRO oncogen, ErbB2 and Grb7, in a breast cancer tissue, normalized against a control gene or genes, and compared to the amount found in a reference breast cancer tissue set, wherein (i) tumors expressing at least one of Bcl1, hepatocyte nuclear factor 3, LIV1, and ER above the mean expression level in the reference tissue set are classified as having a good prognosis for disease free and overall patient survival following surgical removal; (ii) tumors characterized by elevated expression of at least one of lipoprotein lipase, retinol binding protein 4, integrin α7 compared to the reference tissue set are classified as having intermediate prognosis of disease free and overall patient survival following surgical removal; and (iii) tumors expressing either elevated levels of cytokeratins 5 and 17, and GRO oncogen at levels four-fold or greater above the mean expression level in the reference tissue set, or ErbB2 and Grb7 at levels ten-fold or more above the mean expression level in the reference

tissue set are classified as having poor prognosis of disease free and overall patient survival following surgical removal.

Another aspect of the invention is a panel of two or more gene specific primers selected from the group consisting of the forward and reverse primers listed in Table 2.

Yet another aspect of the invention is a method for reverse transcription of a fragmented RNA population in RT-PCR amplification, comprising using a multiplicity of gene specific primers as the reverse primers in the amplification reaction. In a particular embodiment, the method uses between two and about 40,000 gene specific primers in the same amplification reaction. In another embodiment, the gene specific primers are about 18 to 24 bases, such as about 20 bases in length. In another embodiment, the Tm of the primers is about 58-60 °C. The primers can, for example, be selected from the group consisting of the forward and reverse primers listed in Table 2.

The invention also concerns a method of reverse transcriptase driven first strand cDNA synthesis, comprising using a gene specific primer of about 18 to 24 bases in length and having a Tm optimum between about 58 °C and about 60 °C. In a particular embodiment, the first strand cDNA synthesis is followed by PCR DNA amplification, and the primer serves as the reverse primer that drives the PCR amplification. In another embodiment, the method uses a plurality of gene specific primers in the same first strand cDNA synthesis reaction mixture. The number of the gene specific primers can, for example, be between 2 and about 40,000.

In a different aspect, the invention concerns a method of predicting the likelihood of long-term survival of a breast cancer patient without the recurrence of breast cancer, following surgical removal of the primary tumor, comprising determining the expression level of one or more prognostic RNA transcripts or their product in a breast cancer tissue sample obtained from said patient, normalized against the expression level of all RNA transcripts or their products in said breast cancer tissue sample, or of a reference set of RNA transcripts or their products, wherein the prognostic transcript is the transcript of one or more genes selected from the group consisting of: FOXM1, PRAME, Bcl2, STK15, CEGP1, Ki-67, GSTM1, CA9, PR, BBC3, NME1, SURV, GATA3, TFRC, YB-1, DPYD, GSTM3, RPS6KB1, Src, Chk1, ID1, EstR1, p27, CCNB1, XIAP, Chk2, CDC25B, IGF1R, AK055699, P13KC2A, TGFB3, BAGI1, CYP3A4, EpCAM, VEGFC, pS2, hENT1, WISP1, HNF3A, NFKBp65, BRCA2, EGFR, TK1, VDR, Contig51037, pENT1, EPHX1, IF1A, DIABLO, CDH1, HIF1α, IGFBP3, CTSB, and Her2, wherein overexpression of one or more of FOXM1, PRAME, STK15, Ki-67, CA9, NME1, SURV, TFRC, YB-1, RPS6KB1, Src, Chk1, CCNB1, Chk2, CDC25B, CYP3A4, EpCAM, VEGFC, hENT1, BRCA2, EGFR, TK1, VDR, EPHX1, IF1A, Contig51037, CDH1, HIF1α,

IGFBP3, CTSB, Her2, and pENT1 indicates a decreased likelihood of long-term survival without breast cancer recurrence, and the overexpression of one or more of Bcl2, CEGP1, GSTM1, PR, BBC3, GATA3, DPYD, GSTM3, ID1, EstR1, p27, XIAP, IGF1R, AK055699, P13KC2A, TGFB3, BAGI1, pS2, WISP1, HNF3A, NFKBp65, and DIABLO indicates an increased likelihood of long-term survival without breast cancer recurrence.

In a particular embodiment of this method, the expression level of at least 2, preferably at least 5, more preferably at least 10, most preferably at least 15 prognostic transcipts or their expression products is determined.

When the breast cancer is invasive breast carcinoma, including both estrogen receptor (ER) overexpressing (ER positive) and ER negative tumors, the analysis includes determination of the expression levels of the transcripts of at least two of the following genes, or their expression products: FOXM1, PRAME, Bcl2, STK15, CEGP1, Ki-67, GSTM1, PR, BBC3, NME1, SURV, GATA3, TFRC, YB-1, DPYD, Src, CA9, Contig51037, RPS6K1 and Her2.

When the breast cancer is ER positive invasive breast carcinoma, the analysis includes dtermination of the expression levels of the transcripts of at least two of the following genes, or their expression products: PRAME, Bcl2, FOXM1, DIABLO, EPHX1, HIF1A, VEGFC, Ki-67, IGF1R, VDR, NME1, GSTM3, Contig51037, CDC25B, CTSB, p27, CDH1, and IGFBP3.

Just as before, it is preferred to determine the expression levels of at least 5, more preferably at least 10, most preferably at least 15 genes, or their respective expression products.

In a particular embodiment, the expression level of one or more prognostic RNA transcripts is determined, where RNA may, for example, be obtained from a fixed, wax-embedded breast cancer tissue specimen of the patient. The isolation of RNA can, for example, be carried out following any of the procedures described above or throughout the application, or by any other method known in the art.

In yet another aspect, the invention concerns an array comprising polynucleotides hybridizing to the following genes: FOXM1, PRAME, Bcl2, STK15, CEGP1, Ki-67, GSTM1, PR, BBC3, NME1, SURV, GATA3, TFRC, YB-1, DPYD, CA9, Contig51037, RPS6K1 and Her2, immobilized on a solid surface.

In a particular embodiment, the array comprosies polynucleotides hybridizing to the following genes: FOXM1, PRAME, Bcl2, STK15, CEGP1, Ki-67, GSTM1, CA9, PR, BBC3, NME1, SURV, GATA3, TFRC, YB-1, DPYD, GSTM3, RPS6KB1, Src, Chk1, ID1, EstR1, p27, CCNB1, XIAP, Chk2, CDC25B, IGF1R, AK055699, P13KC2A, TGFB3, BAGI1, CYP3A4, EpCAM, VEGFC, pS2, hENT1, WISP1, HNF3A, NFKBp65, BRCA2, EGFR, TK1, VDR, Contig51037, pENT1, EPHX1, IF1A, CDH1, HIF1α, IGFBP3, CTSB, Her2 and DIABLO.

In a further aspect, the invention concerns a method of predicting the likelihood of long-term survival of a patient diagnosed with invasive breast cancer, without the recurrence of breast cancer, following surgical removal of the primary tumor, comprising the steps of:

(1) determining the expression levels of the RNA transcripts or the expression products of genes of a gene set selected from the group consisting of

(a) Bcl2, cyclinG1, NFKBp65, NME1, EPHX1, TOP2B, DR5, TERC, Src, DIABLO;

(b) Ki67, XIAP, hENT1, TS, CD9, p27, cyclinG1, pS2, NFKBp65, CYP3A4;

(c) GSTM1, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, NFKBp65, ErbB3;

(d) PR, NME1, XIAP, upa, cyclinG1, Contig51037, TERC, EPHX1, ALDH1A3, CTSL;

(e) CA9, NME1, TERC, cyclinG1, EPHX1, DPYD, Src, TOP2B, NFKBp65, VEGFC;

(f) TFRC, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, ErbB3, NFKBp65;

(g) Bcl2, PRAME, cyclinG1, FOXM1, NFKBp65, TS, XIAP, Ki67, CYP3A4, p27;

(h) FOXM1, cyclinG1, XIAP, Contig51037, PRAME, TS, Ki67, PDGFRa, p27, NFKBp65;

(i) PRAME, FOXM1, cyclinG1, XIAP, Contig51037, TS, Ki6, PDGFRa, p27, NFKBp65;

(j) Ki67, XIAP, PRAME, hENT1, contig51037, TS, CD9, p27, ErbB3, cyclinG1;

(k) STK15, XIAP, PRAME, PLAUR, p27, CTSL, CD18, PREP, p53, RPS6KB1;

(l) GSTM1, XIAP, PRAME, p27, Contig51037, ErbB3, GSTp, EREG, ID1, PLAUR;

(m) PR, PRAME, NME1, XIAP, PLAUR, cyclinG1, Contig51037, TERC, EPHX1, DR5;

(n) CA9, FOXM1, cyclinG1, XIAP, TS, Ki67, NFKBp65, CYP3A4, GSTM3, p27;

(o) TFRC, XIAP, PRAME, p27, Contig51037, ErbB3, DPYD, TERC, NME1, VEGFC; and

(p) CEGP1, PRAME, hENT1, XIAP, Contig51037, ErbB3, DPYD, NFKBp65, ID1, TS

in a breast cancer tissue sample obtained from said patient, normalized against the expression levels of all RNA transcripts or their products in said breast cancer tissue sample, or of a reference set of RNA transcripts or their products;

(2) subjecting the data obtained in step (a) to statistical analysis; and

(3) determining whether the likelihood of said long-term survival has increased or decreased.

In a still further aspect, the invention concerns a method of predicting the likelihood of long-term survival of a patient diagnosed with estrogen receptor (ER)-positive invasive breast cancer, without the recurrence of breast cancer, following surgical removal of the primary tumor, comprising the steps of:

(1) determining the expression levels of the RNA transcripts or the expression products of genes of a gene set selected from the group consisting of

(a) PRAME, p27, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(b) Contig51037, EPHX1, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;

(c) Bcl2, hENT1, FOXM1, Contig51037, cyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;

(d) HIF1A, PRAME, p27, IGFBP2, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(e) IGF1R, PRAME, EPHX1, Contig51037, cyclinG1, Bcl2, NME1, PTEN, TBP, TIMP2;

(f) FOXM1, Contig51037, VEGFC, TBP, HIF1A, DPYD, RAD51C, DCR3, cyclinG1, BAG1;

(g) EPHX1, Contig51037, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;

(h) Ki67, VEGFC, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;

(i) CDC25B, Contig51037, hENT1, Bcl2, HLAG, TERC, NME1, upa, ID1, CYP;

(j) VEGFC, Ki67, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;

(k) CTSB, PRAME, p27, IGFBP2, EPHX1, CTSL, BAD, DR5, DCR3, XIAP;

(l) DIABLO, Ki67, hENT1, TIMP2, ID1, p27, KRT19, IGFBP2, TS, PDGFB;

(m) p27, PRAME, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(n) CDH1; PRAME, VEGFC; HIF1A; DPYD, TIMP2, CYP3A4, EstR1, RBP4, p27;

(o) IGFBP3, PRAME, p27, Bcl2, XIAP, EstR1, Ki67, TS, Src, VEGF;

(p) GSTM3, PRAME, p27, IGFBP3, XIAP, FGF2, hENT1, PTEN, EstR1, APC;

(q) hENT1, Bcl2, FOXM1, Contig51037, CyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;

(r) STK15, VEGFC, PRAME, p27, GCLC, hENT1, ID1, TIMP2, EstR1, MCP1;

(s) NME1, PRAM, p27, IGFBP3, XIAP, PTEN, hENT1, Bcl2, CYP3A4, HLAG;

(t) VDR, Bcl2, p27, hENT1, p53, PI3KC2A, EIF4E, TFRC, MCM3, ID1;

(u) EIF4E, Contig51037, EPHX1, cyclinG1, Bcl2, DR5, TBP, PTEN, NME1, HER2;

(v) CCNB1, PRAME, VEGFC, HIF1A, hENT1, GCLC, TIMP2, ID1, p27, upa;

(w) ID1, PRAME, DIABLO, hENT1, p27, PDGFRa, NME1, BIN1, BRCA1, TP;

(x). FBXO5, PRAME, IGFBP3, p27, GSTM3, hENT1, XIAP, FGF2, TS, PTEN;

(y) GUS, HIA1A, VEGFC, GSTM3, DPYD, hENT1, EBXO5, CA9, CYP, KRT18; and

(z) Bclx, Bcl2, hENT1, Contig51037, HLAG, CD9, ID1, BRCA1, BIN1, HBEGF;

(2) subjecting the data obtained in step (1) to statistical analysis; and

(3) determining whether the likelihood of said long-term survival has increased or decreased.

In a different aspect, the invention concerns an array comprising polynucleotides hybridizing to a gene set selected from the group consisting of:

(a) Bcl2, cyclinG1, NFKBp65, NME1, EPHX1, TOP2B, DR5, TERC, Src, DIABLO;

(b) Ki67, XIAP, hENT1, TS, CD9, p27, cyclinG1, pS2, NFKBp65, CYP3A4;

(c) GSTM1, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, NFKBp65, ErbB3;

(d) PR, NME1, XIAP, upa, cyclinG1, Contig51037, TERC, EPHX1, ALDH1A3, CTSL;

(e) CA9, NME1, TERC, cyclinG1, EPHX1, DPYD, Src, TOP2B, NFKBp65, VEGFC;

(f) TFRC, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, ErbB3, NFKBp65;

(g) Bcl2, PRAME, cyclinG1, FOXM1, NFKBp65, TS, XIAP, Ki67, CYP3A4, p27;

(h) FOXM1, cyclinG1, XIAP, Contig51037, PRAME, TS, Ki67, PDGFRa, p27, NFKBp65;

(i) PRAME, FOXM1, cyclinG1, XIAP, Contig51037, TS, Ki6, PDGFRa, p27, NFKBp65;

(j) Ki67, XIAP, PRAME, hENT1, contig51037, TS, CD9, p27, ErbB3, cyclinG1;

(k) STK15, XIAP, PRAME, PLAUR, p27, CTSL, CD18, PREP, p53, RPS6KB1;

(l) GSTM1, XIAP, PRAME, p27, Contig51037, ErbB3, GSTp, EREG, ID1, PLAUR;

(m) PR, PRAME, NME1, XIAP, PLAUR, cyclinG1, Contig51037, TERC, EPHX1, DR5;

(n) CA9, FOXM1, cyclinG1, XIAP, TS, Ki67, NFKBp65, CYP3A4, GSTM3, p27;

(o) TFRC, XIAP, PRAME, p27, Contig51037, ErbB3, DPYD, TERC, NME1, VEGFC; and

(p) CEGP1, PRAME, hENT1, XIAP, Contig51037, ErbB3, DPYD, NFKBp65, ID1, TS,

immobilized on a solid surface.

In an additional aspect, the invention concerns an array comprising polynucleotides hybridizing to a gene set selected from the group consisting of:

(a) PRAME, p27, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(b) Contig51037, EPHX1, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;

(c) Bcl2, hENT1, FOXM1, Contig51037, cyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;

(d) HIF1A, PRAME, p27, IGFBP2, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(e) IGF1R, PRAME, EPHX1, Contig51037, cyclinG1, Bcl2, NME1, PTEN, TBP, TIMP2;

(f) FOXM1, Contig51037, VEGFC, TBP, HIF1A, DPYD, RAD51C, DCR3, cyclinG1, BAG1;

(g) EPHX1, Contig51037, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;

(h) Ki67, VEGFC, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;

(i) CDC25B, Contig51037, hENT1, Bcl2, HLAG, TERC, NME1, upa, ID1, CYP;

(j) VEGFC, Ki67, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;

(k) CTSB, PRAME, p27, IGFBP2, EPHX1, CTSL, BAD, DR5, DCR3, XIAP;

(l) DIABLO, Ki67, hENT1, TIMP2, ID1, p27, KRT19, IGFBP2, TS, PDGFB;

(m) p27, PRAME, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(n) CDH1; PRAME, VEGFC; HIF1A; DPYD, TIMP2, CYP3A4, EstR1, RBP4, p27;

(o) IGFBP3, PRAME, p27, Bcl2, XIAP, EstR1, Ki67, TS, Src, VEGF;

(p) GSTM3, PRAME, p27, IGFBP3, XIAP, FGF2, hENT1, PTEN, EstR1, APC;

(q) hENT1, Bcl2, FOXM1, Contig51037, CyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;

(r) STK15, VEGFC, PRAME, p27, GCLC, hENT1, ID1, TIMP2, EstR1, MCP1;

(s) NME1, PRAM, p27, IGFBP3, XIAP, PTEN, hENT1, Bcl2, CYP3A4, HLAG;

(t) VDR, Bcl2, p27, hENT1, p53, PI3KC2A, EIF4E, TFRC, MCM3, ID1;

(u) EIF4E, Contig51037, EPHX1, cyclinG1, Bcl2, DR5, TBP, PTEN, NME1, HER2;

(v) CCNB1, PRAME, VEGFC, HIF1A, hENT1, GCLC, TIMP2, ID1, p27, upa;

(w) ID1, PRAME, DIABLO, hENT1, p27, PDGFRa, NME1, BIN1, BRCA1, TP;

(x) FBXO5, PRAME, IGFBP3, p27, GSTM3, hENT1, XIAP, FGF2, TS, PTEN;

(y) GUS, HIA1A, VEGFC, GSTM3, DPYD, hENT1, FBXO5, CA9, CYP, KRT18; and

(z) Bclx, Bcl2, hENT1, Contig51037, HLAG, CD9, ID1, BRCA1, BIN1, HBEGF,

immobilized on a solid surface.

In all aspects, the polynucleotides can be cDNAs ("cDNA arrays") that are typically about 500 to 5000 bases long, although shorter or longer cDNAs can also be used and are within the scope of this invention. Alternatively, the polynucleotids can be oligonucleotides (DNA microarrays), which are typically about 20 to 80 bases long, although shorter and longer oligonucleotides are also suitable and are within the scope of the invention. The solid surface can, for example, be glass or nylon, or any other solid surface typically used in preparing arrays, such as microarrays, and is typically glass.

Brief Description of the Drawings

Figure 1 is a chart illustrating the overall workflow of the process of the invention for measurement of gene expression. In the Figure, FPET stands for "fixed paraffin-embedded tissue," and "RT-PCR" stands for "reverse transcriptase PCR." RNA concentration is determined by using the commercial RiboGreen™ RNA Quantitation Reagent and Protocol.

Figure 2 is a flow chart showing the steps of an RNA extraction method according to the invention alongside a flow chart of a representative commercial method.

Figure 3 is a scheme illustrating the steps of an improved method for preparing fragmented mRNA for expression profiling analysis.

Figure 4 illustrates methods for amplification of RNA prior to RT-PCR.

Figure 5 illustrates an alternative scheme for repair and amplification of fragmented mRNA.

Figure 6 shows the measurement of estrogen receptor mRNA levels in 40 FPE breast cancer specimens via RT-PCR. Three 10 micron sections were used for each measurement. Each data point represents the average of triplicate measurements.

Figure 7 shows the results of the measurement of progesterone receptor mRNA levels in 40 FPE breast cancer specimens via RT-PCR performed as described in the legend of Figure 6 above.

Figure 8 shows results from an IVT/RT-PCR experiment.

Figure 9 is a representation of the expression of 92 genes across 70 FPE breast cancer specimens. The y-axis shows expression as cycle threshold times. These genes are a subset of the genes listed in Table 1.

Table 1 shows a breast cancer gene list.

Table 2 sets forth amplicon and primer sequences used for amplification of fragmented mRNA.

Table 3 shows the Accession Nos. and SEQ ID NOS of the breast cance genes examined.

Detailed Description of the Preferred Embodiment

A. Definitions

Unless defined otherwise, technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Singleton *et al.*, Dictionary of Microbiology and Molecular Biology 2nd ed., J. Wiley & Sons (New York, NY 1994), and March, Advanced Organic Chemistry Reactions, Mechanisms

and Structure 4th ed., John Wiley & Sons (New York, NY 1992), provide one skilled in the art with a general guide to many of the terms used in the present application.

One skilled in the art will recognize many methods and materials similar or equivalent to those described herein, which could be used in the practice of the present invention. Indeed, the present invention is in no way limited to the methods and materials described. For purposes of the present invention, the following terms are defined below.

The term "microarray" refers to an ordered arrangement of hybridizable array elements, preferably polynucleotide probes, on a substrate.

The term "polynucleotide," when used in singular or plural, generally refers to any polyribonucleotide or polydeoxribonucleotide, which may be unmodified RNA or DNA or modified RNA or DNA. Thus, for instance, polynucleotides as defined herein include, without limitation, single- and double-stranded DNA, DNA including single- and double-stranded regions, single- and double-stranded RNA, and RNA including single- and double-stranded regions, hybrid molecules comprising DNA and RNA that may be single-stranded or, more typically, double-stranded or include single- and double-stranded regions. In addition, the term "polynucleotide" as used herein refers to triple-stranded regions comprising RNA or DNA or both RNA and DNA. The strands in such regions may be from the same molecule or from different molecules. The regions may include all of one or more of the molecules, but more typically involve only a region of some of the molecules. One of the molecules of a triple-helical region often is an oligonucleotide. The term "polynucleotide" specifically includes DNAs and RNAs that contain one or more modified bases. Thus, DNAs or RNAs with backbones modified for stability or for other reasons are "polynucleotides" as that term is intended herein. Moreover, DNAs or RNAs comprising unusual bases, such as inosine, or modified bases, such as tritiated bases, are included within the term "polynucleotides" as defined herein. In general, the term "polynucleotide" embraces all chemically, enzymatically and/or metabolically modified forms of unmodified polynucleotides, as well as the chemical forms of DNA and RNA characteristic of viruses and cells, including simple and complex cells.

The term "oligonucleotide" refers to a relatively short polynucleotide, including, without limitation, single-stranded deoxyribonucleotides, single- or double-stranded ribonucleotides, RNA:DNA hybrids and double-stranded DNAs. Oligonucleotides, such as single-stranded DNA probe oligonucleotides, are often synthesized by chemical methods, for example using automated oligonucleotide synthesizers that are commercially available. However, oligonucleotides can be made by a variety of other methods, including *in vitro* recombinant DNA-mediated techniques and by expression of DNAs in cells and organisms.

The terms "differentially expressed gene," "differential gene expression" and their synonyms, which are used interchangeably, refer to a gene whose expression is activated to a higher or lower level in a subject suffering from a disease, specifically cancer, such as breast cancer, relative to its expression in a normal or control subject. The terms also include genes whose expression is activated to a higher or lower level at different stages of the same disease. It is also understood that a differentially expressed gene may be either activated or inhibited at the nucleic acid level or protein level, or may be subject to alternative splicing to result in a different polypeptide product. Such differences may be evidenced by a change in mRNA levels, surface expression, secretion or other partitioning of a polypeptide, for example. Differential gene expression may include a comparison of expression between two or more genes, or a comparison of the ratios of the expression between two or more genes, or even a comparison of two differently processed products of the same gene, which differ between normal subjects and subjects suffering from a disease, specifically cancer, or between various stages of the same disease. Differential expression includes both quantitative, as well as qualitative, differences in the temporal or cellular expression pattern in a gene or its expression products among, for example, normal and diseased cells, or among cells which have undergone different disease events or disease stages. For the purpose of this invention, "differential gene expression" is considered to be present when there is at least an about two-fold, preferably at least about four-fold, more preferably at least about six-fold, most preferably at least about ten-fold difference between the expression of a given gene in normal and diseased subjects, or in various stages of disease development in a diseased subject.

The phrase "gene amplification" refers to a process by which multiple copies of a gene or gene fragment are formed in a particular cell or cell line. The duplicated region (a stretch of amplified DNA) is often referred to as "amplicon." Usually, the amount of the messenger RNA (mRNA) produced, *i.e.*, the level of gene expression, also increases in the proportion of the number of copies made of the particular gene expressed.

The term "prognosis" is used herein to refer to the prediction of the likelihood of cancer-attributable death or progression, including recurrence, metastatic spread, and drug resistance, of a neoplastic disease, such as breast cancer. The term "prediction" is used herein to refer to the likelihood that a patient will respond either favorably or unfavorably to a drug or set of drugs, and also the extent of those responses. The predictive methods of the present invention can be used clinically to make treatment decisions by choosing the most appropriate treatment modalities for any particular patient. The predictive methods of the present invention are valuable tools in predicting if a patient is likely to respond favorably to a treatment regimen, such

as surgical intervention, chemotherapy with a given drug or drug combination, and/or radiation therapy.

The term "increased resistance" to a particular drug or treatment option, when used in accordance with the present invention, means decreased response to a standard dose of the drug or to a standard treatment protocol.

The term "decreased sensitivity" to a particular drug or treatment option, when used in accordance with the present invention, means decreased response to a standard dose of the drug or to a standard treatment protocol, where decreased response can be compensated for (at least partially) by increasing the dose of drug, or the intensity of treatment.

"Patient response" can be assessed using any endpoint indicating a benefit to the patient, including, without limitation, (1) inhibition, to some extent, of tumor growth, including slowing down and complete growth arrest; (2) reduction in the number of tumor cells; (3) reduction in tumor size; (4) inhibition (i.e., reduction, slowing down or complete stopping) of tumor cell infiltration into adjacent peripheral organs and/or tissues; (5) inhibition (i.e. reduction, slowing down or complete stopping) of metastasis; (6) enhancement of anti-tumor immune response, which may, but does not have to, result in the regression or rejection of the tumor; (7) relief, to some extent, of one or more symptoms associated with the tumor; (8) increase in the length of survival following treatment; and/or (9) decreased mortality at a given point of time following treatment.

The term "treatment" refers to both therapeutic treatment and prophylactic or preventative measures, wherein the object is to prevent or slow down (lessen) the targeted pathologic condition or disorder. Those in need of treatment include those already with the disorder as well as those prone to have the disorder or those in whom the disorder is to be prevented. In tumor (*e.g.*, cancer) treatment, a therapeutic agent may directly decrease the pathology of tumor cells, or render the tumor cells more susceptible to treatment by other therapeutic agents, *e.g.*, radiation and/or chemotherapy.

The term "tumor," as used herein, refers to all neoplastic cell growth and proliferation, whether malignant or benign, and all pre-cancerous and cancerous cells and tissues.

The terms "cancer" and "cancerous" refer to or describe the physiological condition in mammals that is typically characterized by unregulated cell growth. Examples of cancer include but are not limited to, breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer, pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, and brain cancer.

The "pathology" of cancer includes all phenomena that compromise the well-being of the patient. This includes, without limitation, abnormal or uncontrollable cell growth, metastasis, interference with the normal functioning of neighboring cells, release of cytokines or other secretory products at abnormal levels, suppression or aggravation of inflammatory or immunological response, neoplasia, premalignancy, malignancy, invasion of surrounding or distant tissues or organs, such as lymph nodes, etc.

"Stringency" of hybridization reactions is readily determinable by one of ordinary skill in the art, and generally is an empirical calculation dependent upon probe length, washing temperature, and salt concentration. In general, longer probes require higher temperatures for proper annealing, while shorter probes need lower temperatures. Hybridization generally depends on the ability of denatured DNA to reanneal when complementary strands are present in an environment below their melting temperature. The higher the degree of desired homology between the probe and hybridizable sequence, the higher the relative temperature which can be used. As a result, it follows that higher relative temperatures would tend to make the reaction conditions more stringent, while lower temperatures less so. For additional details and explanation of stringency of hybridization reactions, see Ausubel et al., Current Protocols in Molecular Biology, Wiley Interscience Publishers, (1995).

"Stringent conditions" or "high stringency conditions", as defined herein, typically: (1) employ low ionic strength and high temperature for washing, for example 0.015 M sodium chloride/0.0015 M sodium citrate/0.1% sodium dodecyl sulfate at 50°C; (2) employ during hybridization a denaturing agent, such as formamide, for example, 50% (v/v) formamide with 0.1% bovine serum albumin/0.1% Ficoll/0.1% polyvinylpyrrolidone/50mM sodium phosphate buffer at pH 6.5 with 750 mM sodium chloride, 75 mM sodium citrate at 42°C; or (3) employ 50% formamide, 5 x SSC (0.75 M NaCl, 0.075 M sodium citrate), 50 mM sodium phosphate (pH 6.8), 0.1% sodium pyrophosphate, 5 x Denhardt's solution, sonicated salmon sperm DNA (50 µg/ml), 0.1% SDS, and 10% dextran sulfate at 42°C, with washes at 42°C in 0.2 x SSC (sodium chloride/sodium citrate) and 50% formamide at 55°C, followed by a high-stringency wash consisting of 0.1 x SSC containing EDTA at 55°C.

"Moderately stringent conditions" may be identified as described by Sambrook et al., Molecular Cloning: A Laboratory Manual, New York: Cold Spring Harbor Press, 1989, and include the use of washing solution and hybridization conditions (e.g., temperature, ionic strength and %SDS) less stringent that those described above. An example of moderately stringent conditions is overnight incubation at 37°C in a solution comprising: 20% formamide, 5 x SSC (150 mM NaCl, 15 mM trisodium citrate), 50 mM sodium phosphate (pH 7.6), 5 x

Denhardt's solution, 10% dextran sulfate, and 20 mg/ml denatured sheared salmon sperm DNA, followed by washing the filters in 1 x SSC at about 37-50°C. The skilled artisan will recognize how to adjust the temperature, ionic strength, etc. as necessary to accommodate factors such as probe length and the like.

In the context of the present invention, reference to "at least one," "at least two," "at least five," etc. of the genes listed in any particular gene set means any one or any and all combinations of the genes listed.

The terms "splicing" and "RNA splicing" are used interchangeably and refer to RNA processing that removes introns and joins exons to produce mature mRNA with continuous coding sequence that moves into the cytoplasm of an eukaryotic cell.

In theory, the term "exon" refers to any segment of an interrupted gene that is represented in the mature RNA product (B. Lewin. *Genes IV* Cell Press, Cambridge Mass. 1990). In theory the term "intron" refers to any segment of DNA that is transcribed but removed from within the transcript by splicing together the exons on either side of it. Operationally, exon sequences occur in the mRNA sequence of a gene as defined by Ref. Seq ID numbers. Operationally, intron sequences are the intervening sequences within the genomic DNA of a gene, bracketed by exon sequences and having GT and AG splice consensus sequences at their 5' and 3' boundaries.

B. Detailed Description

The practice of the present invention will employ, unless otherwise indicated, conventional techniques of molecular biology (including recombinant techniques), microbiology, cell biology, and biochemistry, which are within the skill of the art. Such techniques are explained fully in the literature, such as, "Molecular Cloning: A Laboratory Manual", 2nd edition (Sambrook et al., 1989); "Oligonucleotide Synthesis" (M.J. Gait, ed., 1984); "Animal Cell Culture" (R.I. Freshney, ed., 1987); "Methods in Enzymology" (Academic Press, Inc.); "Handbook of Experimental Immunology", 4th edition (D.M. Weir & C.C. Blackwell, eds., Blackwell Science Inc., 1987); "Gene Transfer Vectors for Mammalian Cells" (J.M. Miller & M.P. Calos, eds., 1987); "Current Protocols in Molecular Biology" (F.M. Ausubel et al., eds., 1987); and "PCR: The Polymerase Chain Reaction", (Mullis et al., eds., 1994).

1. *Gene Expression Profiling*

In general, methods of gene expression profiling can be divided into two large groups: methods based on hybridization analysis of polynucleotides, and methods based on sequencing of polynucleotides. The most commonly used methods known in the art for the quantification of mRNA expression in a sample include northern blotting and *in situ* hybridization (Parker & Barnes, *Methods in Molecular Biology* 106:247-283 (1999)); RNAse protection assays (Hod,

Biotechniques 13:852-854 (1992)); and reverse transcription polymerase chain reaction (RT-PCR) (Weis *et al., Trends in Genetics* 8:263-264 (1992)). Alternatively, antibodies may be employed that can recognize specific duplexes, including DNA duplexes, RNA duplexes, and DNA-RNA hybrid duplexes or DNA-protein duplexes. Representative methods for sequencing-based gene expression analysis include Serial Analysis of Gene Expression (SAGE), and gene expression analysis by massively parallel signature sequencing (MPSS).

2. *Reverse Transcriptase PCR (RT-PCR)*

Of the techniques listed above, the most sensitive and most flexible quantitative method is RT-PCR, which can be used to compare mRNA levels in different sample populations, in normal and tumor tissues, with or without drug treatment, to characterize patterns of gene expression, to discriminate between closely related mRNAs, and to analyze RNA structure.

The first step is the isolation of mRNA from a target sample. The starting material is typically total RNA isolated from human tumors or tumor cell lines, and corresponding normal tissues or cell lines, respectively. Thus RNA can be isolated from a variety of primary tumors, including breast, lung, colon, prostate, brain, liver, kidney, pancreas, spleen, thymus, testis, ovary, uterus, etc., tumor, or tumor cell lines, with pooled DNA from healthy donors. If the source of mRNA is a primary tumor, mRNA can be extracted, for example, from frozen or archived paraffin-embedded and fixed (e.g. formalin-fixed) tissue samples.

General methods for mRNA extraction are well known in the art and are disclosed in standard textbooks of molecular biology, including Ausubel *et al.*, Current Protocols of Molecular Biology, John Wiley and Sons (1997). Methods for RNA extraction from paraffin embedded tissues are disclosed, for example, in Rupp and Locker, *Lab Invest.* 56:A67 (1987), and De Andrés *et al., BioTechniques* 18:42044 (1995). In particular, RNA isolation can be performed using purification kit, buffer set and protease from commercial manufacturers, such as Qiagen, according to the manufacturer's instructions. For example, total RNA from cells in culture can be isolated using Qiagen RNeasy mini-columns. Other commercially available RNA isolation kits include MasterPure™ Complete DNA and RNA Purification Kit (EPICENTRE®, Madison, WI), and Paraffin Block RNA Isolation Kit (Ambion, Inc.). Total RNA from tissue samples can be isolated using RNA Stat-60 (Tel-Test). RNA prepared from tumor can be isolated, for example, by cesium chloride density gradient centrifugation.

As RNA cannot serve as a template for PCR, the first step in gene expression profiling by RT-PCR is the reverse transcription of the RNA template into cDNA, followed by its exponential amplification in a PCR reaction. The two most commonly used reverse transcriptases are avilo myeloblastosis virus reverse transcriptase (AMV-RT) and Moloney murine leukemia virus

reverse transcriptase (MMLV-RT). The reverse transcription step is typically primed using specific primers, random hexamers, or oligo-dT primers, depending on the circumstances and the goal of expression profiling. For example, extracted RNA can be reverse-transcribed using a GeneAmp RNA PCR kit (Perkin Elmer, CA, USA), following the manufacturer's instructions. The derived cDNA can then be used as a template in the subsequent PCR reaction.

Although the PCR step can use a variety of thermostable DNA-dependent DNA polymerases, it typically employs the Taq DNA polymerase, which has a 5'-3' nuclease activity but lacks a 3'-5' proofreading endonuclease activity. Thus, TaqMan® PCR typically utilizes the 5'-nuclease activity of Taq or Tth polymerase to hydrolyze a hybridization probe bound to its target amplicon, but any enzyme with equivalent 5' nuclease activity can be used. Two oligonucleotide primers are used to generate an amplicon typical of a PCR reaction. A third oligonucleotide, or probe, is designed to detect nucleotide sequence located between the two PCR primers. The probe is non-extendible by Taq DNA polymerase enzyme, and is labeled with a reporter fluorescent dye and a quencher fluorescent dye. Any laser-induced emission from the reporter dye is quenched by the quenching dye when the two dyes are located close together as they are on the probe. During the amplification reaction, the Taq DNA polymerase enzyme cleaves the probe in a template-dependent manner. The resultant probe fragments disassociate in solution, and signal from the released reporter dye is free from the quenching effect of the second fluorophore. One molecule of reporter dye is liberated for each new molecule synthesized, and detection of the unquenched reporter dye provides the basis for quantitative interpretation of the data.

TaqMan® RT-PCR can be performed using commercially available equipment, such as, for example, ABI PRISM 7700™ Sequence Detection System™ (Perkin-Elmer-Applied Biosystems, Foster City, CA, USA), or Lightcycler (Roche Molecular Biochemicals, Mannheim, Germany). In a preferred embodiment, the 5' nuclease procedure is run on a real-time quantitative PCR device such as the ABI PRISM 7700™ Sequence Detection System™. The system consists of a thermocycler, laser, charge-coupled device (CCD), camera and computer. The system amplifies samples in a 96-well format on a thermocycler. During amplification, laser-induced fluorescent signal is collected in real-time through fiber optics cables for all 96 wells, and detected at the CCD. The system includes software for running the instrument and for analyzing the data.

5'-Nuclease assay data are initially expressed as Ct, or the threshold cycle. As discussed above, fluorescence values are recorded during every cycle and represent the amount of product

amplified to that point in the amplification reaction. The point when the fluorescent signal is first recorded as statistically significant is the threshold cycle (C_t).

To minimize errors and the effect of sample-to-sample variation, RT-PCR is usually performed using an internal standard. The ideal internal standard is expressed at a constant level among different tissues, and is unaffected by the experimental treatment. RNAs most frequently used to normalize patterns of gene expression are mRNAs for the housekeeping genes glyceraldehyde-3-phosphate-dehydrogenase (GAPDH) and β-actin.

A more recent variation of the RT-PCR technique is the real time quantitative PCR, which measures PCR product accumulation through a dual-labeled fluorigenic probe (i.e., TaqMan® probe). Real time PCR is compatible both with quantitative competitive PCR, where internal competitor for each target sequence is used for normalization, and with quantitative comparative PCR using a normalization gene contained within the sample, or a housekeeping gene for RT-PCR. For further details see, e.g. Held *et al., Genome Research* 6:986-994 (1996).

3. *Microarrays*

Differential gene expression can also be identified, or confirmed using the microarray technique. Thus, the expression profile of breast cancer-associated genes can be measured in either fresh or paraffin-embedded tumor tissue, using microarray technology. In this method, polynucleotide sequences of interest are plated, or arrayed, on a microchip substrate. The arrayed sequences are then hybridized with specific DNA probes from cells or tissues of interest. Just as in the RT-PCR method, the source of mRNA typically is total RNA isolated from human tumors or tumor cell lines, and corresponding normal tissues or cell lines. Thus RNA can be isolated from a variety of primary tumors or tumor cell lines. If the source of mRNA is a primary tumor, mRNA can be extracted, for example, from frozen or archived paraffin-embedded and fixed (e.g. formalin-fixed) tissue samples, which are routinely prepared and preserved in everyday clinical practice.

In a specific embodiment of the microarray technique, PCR amplified inserts of cDNA clones are applied to a substrate in a dense array. Preferably at least 10,000 nucleotide sequences are applied to the substrate. The microarrayed genes, immobilized on the microchip at 10,000 elements each, are suitable for hybridization under stringent conditions. Fluorescently labeled cDNA probes may be generated through incorporation of fluorescent nucleotides by reverse transcription of RNA extracted from tissues of interest. Labeled cDNA probes applied to the chip hybridize with specificity to each spot of DNA on the array. After stringent washing to remove non-specifically bound probes, the chip is scanned by confocal laser microscopy or by another detection method, such as a CCD camera. Quantitation of hybridization of each arrayed

element allows for assessment of corresponding mRNA abundance. With dual color fluorescence, separately labeled cDNA probes generated from two sources of RNA are hybridized pairwise to the array. The relative abundance of the transcripts from the two sources corresponding to each specified gene is thus determined simultaneously. The miniaturized scale of the hybridization affords a convenient and rapid evaluation of the expression pattern for large numbers of genes. Such methods have been shown to have the sensitivity required to detect rare transcripts, which are expressed at a few copies per cell, and to reproducibly detect at least approximately two-fold differences in the expression levels (Schena *et al., Proc. Natl. Acad. Sci. USA* 93(2):106-149 (1996)). Microarray analysis can be performed by commercially available equipment, following manufacturer's protocols, such as by using the Affymetrix GenChip technology, or Incyte's microarray technology.

The development of microarray methods for large-scale analysis of gene expression makes it possible to search systematically for molecular markers of cancer classification and outcome prediction in a variety of tumor types.

4. *Serial Analysis of Gene Expression (SAGE)*

Serial analysis of gene expression (SAGE) is a method that allows the simultaneous and quantitative analysis of a large number of gene transcripts, without the need of providing an individual hybridization probe for each transcript. First, a short sequence tag (about 10-14 bp) is generated that contains sufficient information to uniquely identify a transcript, provided that the tag is obtained from a unique position within each transcript. Then, many transcripts are linked together to form long serial molecules, that can be sequenced, revealing the identity of the multiple tags simultaneously. The expression pattern of any population of transcripts can be quantitatively evaluated by determining the abundance of individual tags, and identifying the gene corresponding to each tag. For more details see, e.g. Velculescu *et al., Science* 270:484-487 (1995); and Velculescu *et al., Cell* 88:243-51 (1997).

5. *Gene Expression Analysis by Massively Parallel Signature Sequencing (MPSS)*

This method, described by Brenner *et al., Nature Biotechnology* 18:630-634 (2000), is a sequencing approach that combines non-gel-based signature sequencing with *in vitro* cloning of millions of templates on separate 5 μm diameter microbeads. First, a microbead library of DNA templates is constructed by *in vitro* cloning. This is followed by the assembly of a planar array of the template-containing microbeads in a flow cell at a high density (typically greater than 3×10^6 microbeads/cm^2). The free ends of the cloned templates on each microbead are analyzed simultaneously, using a fluorescence-based signature sequencing method that does not require DNA fragment separation. This method has been shown to simultaneously and accurately

provide, in a single operation, hundreds of thousands of gene signature sequences from a yeast cDNA library.

6. *General Description of the mRNA Isolation, Purification and Amplification Methods of the Invention*

The steps of a representative protocol of the invention, including mRNA isolation, purification, primer extension and amplification are illustrated in Figure 1. As shown in Figure 1, this representative process starts with cutting about 10 μm thick sections of paraffin-embedded tumor tissue samples. The RNA is then extracted, and protein and DNA are removed, following the method of the invention described below. After analysis of the RNA concentration, RNA repair and/or amplification steps may be included, if necessary, and RNA is reverse transcribed using gene specific promoters followed by RT-PCR. Finally, the data are analyzed to identify the best treatment option(s) available to the patient on the basis of the characteristic gene expression pattern identified in the tumor sample examined. The individual steps of this protocol will be discussed in greater detail below.

7. *Improved Method for Isolation of Nucleic Acid from Archived Tissue Specimens*

As discussed above, in the first step of the method of the invention, total RNA is extracted from the source material of interest, including fixed, paraffin-embedded tissue specimens, and purified sufficiently to act as a substrate in an enzyme assay. Despite the availability of commercial products, and the extensive knowledge available concerning the isolation of nucleic acid, such as RNA, from tissues, isolation of nucleic acid (RNA) from fixed, paraffin-embedded tissue specimens (FPET) is not without difficulty.

In one aspect, the present invention concerns an improved method for the isolation of nucleic acid from archived, e.g. FPET tissue specimens. Measured levels of mRNA species are useful for defining the physiological or pathological status of cells and tissues. RT-PCR (which is discussed above) is one of the most sensitive, reproducible and quantitative methods for this "gene expression profiling". Paraffin-embedded, formalin-fixed tissue is the most widely available material for such studies. Several laboratories have demonstrated that it is possible to successfully use fixed-paraffin-embedded tissue (FPET) as a source of RNA for RT-PCR (Stanta *et al., Biotechniques* 11:304-308 (1991); Stanta *et al., Methods Mol. Biol.* 86:23-26 (1998); Jackson *et al., Lancet* 1:1391 (1989); Jackson *et al., J. Clin. Pathol.* 43:499-504 (1999); Finke *et al., Biotechniques* 14:448-453 (1993); Goldsworthy *et al., Mol. Carcinog.* 25:86-91 (1999); Stanta and Bonin, *Biotechniques* 24:271-276 (1998); Godfrey *et al., J. Mol. Diagnostics* 2:84 (2000); Specht *et al., J. Mol. Med.* 78:B27 (2000); Specht *et al., Am. J. Pathol.* 158:419-429 (2001)). This allows gene expression profiling to be carried out on the most commonly available

source of human biopsy specimens, and therefore potentially to create new valuable diagnostic and therapeutic information.

The most widely used protocols utilize hazardous organic solvents, such as xylene, or octane (Finke *et al., supra*) to dewax the tissue in the paraffin blocks before nucleic acid (RNA and/or DNA) extraction. Obligatory organic solvent removal (e.g. with ethanol) and rehydration steps follow, which necessitate multiple manipulations, and addition of substantial total time to the protocol, which can take up to several days. Commercial kits and protocols for RNA extraction from FPET [MasterPure™ Complete DNA and RNA Purification Kit (EPICENTRE®, Madison, WI); Paraffin Block RNA Isolation Kit (Ambion, Inc.) and RNeasy™ Mini kit (Qiagen, Chatsworth, CA)] use xylene for deparaffinization, in procedures which typically require multiple centrifugations and ethanol buffer changes, and incubations following incubation with xylene.

The present invention provides an improved nucleic acid extraction protocol that produces nucleic acid, in particular RNA, sufficiently intact for gene expression measurements. The key step in the nucleic acid extraction protocol herein is the performance of dewaxing without the use of any organic solvent, thereby eliminating the need for multiple manipulations associated with the removal of the organic solvent, and substantially reducing the total time to the protocol. According to the invention, wax, e.g. paraffin is removed from wax-embedded tissue samples by incubation at 65-75 °C in a lysis buffer that solubilizes the tissue and hydrolyzes the protein, following by cooling to solidify the wax.

Figure 2 shows a flow chart of an RNA extraction protocol of the present invention in comparison with a representative commercial method, using xylene to remove wax. The times required for individual steps in the processes and for the overall processes are shown in the chart. As shown, the commercial process requires approximately 50% more time than the process of the invention.

The lysis buffer can be any buffer known for cell lysis. It is, however, preferred that oligo-dT-based methods of selectively purifying polyadenylated mRNA not be used to isolate RNA for the present invention, since the bulk of the mRNA molecules are expected to be fragmented and therefore will not have an intact polyadenylated tail, and will not be recovered or available for subsequent analytical assays. Otherwise, any number of standard nucleic acid purification schemes can be used. These include chaotrope and organic solvent extractions, extraction using glass beads or filters, salting out and precipitation based methods, or any of the purification methods known in the art to recover total RNA or total nucleic acids from a biological source.

Lysis buffers are commercially available, such as, for example, from Qiagen, Epicentre, or Ambion. A preferred group of lysis buffers typically contains urea, and Proteinase K or other protease. Proteinase K is very useful in the isolation of high quality, undamaged DNA or RNA, since most mammalian DNases and RNases are rapidly inactivated by this enzyme, especially in the presence of 0.5 - 1% sodium dodecyl sulfate (SDS). This is particularly important in the case of RNA, which is more susceptible to degradation than DNA. While DNases require metal ions for activity, and can therefore be easily inactivated by chelating agents, such as EDTA, there is no similar co-factor requirement for RNases.

Cooling and resultant solidification of the wax permits easy separation of the wax from the total nucleic acid, which can be conveniently precipitated, e.g. by isopropanol. Further processing depends on the intended purpose. If the proposed method of RNA analysis is subject to bias by contaminating DNA in an extract, the RNA extract can be further treated, e.g. by DNase, post purification to specifically remove DNA while preserving RNA. For example, if the goal is to isolate high quality RNA for subsequent RT-PCR amplification, nucleic acid precipitation is followed by the removal of DNA, usually by DNase treatment. However, DNA can be removed at various stages of nucleic acid isolation, by DNase or other techniques well known in the art.

While the advantages of the nucleic acid extraction protocol of the invention are most apparent for the isolation of RNA from archived, paraffin embedded tissue samples, the wax removal step of the present invention, which does not involve the use of an organic solvent, can also be included in any conventional protocol for the extraction of total nucleic acid (RNA and DNA) or DNA only. All of these aspects are specifically within the scope of the invention.

By using heat followed by cooling to remove paraffin, the process of the present invention saves valuable processing time, and eliminates a series of manipulations, thereby potentially increasing the yield of nucleic acid. Indeed, experimental evidence presented in the examples below, demonstrates that the method of the present invention does not compromise RNA yield.

8. *5'-multiplexed Gene Specific Priming of Reverse Transcription*

RT-PCR requires reverse transcription of the test RNA population as a first step. The most commonly used primer for reverse transcription is oligo-dT, which works well when RNA is intact. However, this primer will not be effective when RNA is highly fragmented as is the case in FPE tissues.

The present invention includes the use of gene specific primers, which are roughly 20 bases in length with a Tm optimum between about 58 °C and 60 °C. These primers will also serve as the reverse primers that drive PCR DNA amplification.

Another aspect of the invention is the inclusion of multiple gene-specific primers in the same reaction mixture. The number of such different primers can vary greatly and can be as low as two and as high as 40,000 or more. Table 2 displays examples of reverse primers that can be successfully used in carrying out the methods of the invention. Figure 9 shows expression data obtained using this multiplexed gene-specific priming strategy. Specifically, Figure 9 is a representation of the expression of 92 genes (a subset of genes listed in Table 1) across 70 FPE breast cancer specimens. The y-axis shows expression as cycle threshold times.

An alternative approach is based on the use of random hexamers as primers for cDNA synthesis. However, we have experimentally demonstrated that the method of using a multiplicity of gene-specific primers is superior over the known approach using random hexamers.

9. *Preparation of Fragmented mRNA for Expression Profiling Assays*

It is of interest to analyze the abundance of specific mRNA species in biological samples, since this expression profile provides an index of the physiological state of that sample. mRNA is notoriously difficult to extract and maintain in its native state, consequently, mRNA recovered from biological sources is often fragmented or somewhat degraded. This is especially true of human tissue specimen which have been chemically fixed and stored for extended periods of time.

In one aspect, the present invention provides a means of preparing the mRNA extracted from various sources, including archived tissue specimens, for expression profiling in a way that its relative abundance is preserved and the mRNA's of interest can be successfully measured. This method is useful as a means of preparing mRNA for analysis by any of the known expression profiling methods, including RT-PCR coupled with 5' exonuclease of reporter probes (TaqMan® type assays), as discussed above, flap endonuclease assays (Cleavase® and Invader® type assays), oligonucleotide hybridization arrays, cDNA hybridization arrays, oligonucleotide ligation assays, 3' single nucleotide extension assays and other assays designed to assess the abundance of specific mRNA sequences in a biological sample.

According to the method of the invention, total RNA is extracted from the source material and sufficiently purified to act as a substrate in an enzyme assay. The extraction procedure, including a new and improved way of removing the wax (e.g. paraffin) used for embedding the tissue samples, has been discussed above. It has also been noted that it is preferred that oligo-dT based methods of selectively purifying polyadenylated mRNA not be used to isolate RNA for this invention since the bulk of the mRNA is expected to be fragmented, will not be polyadenylated

and, therefore, will not be recovered and available for subsequent analytical assays if an oligo-dT based method is used.

A diagram of an improved method for repairing fragmented RNA is shown in Figure 3. The fragmented RNA purified from the tissue sample is mixed with universal or gene-specific, single-stranded, DNA templates for each mRNA species of interest. These templates may be full length DNA copies of the mRNA derived from cloned gene sources, they may be fragments of the gene representing only the segment of the gene to be assayed, they may be a series of long oligonucleotides representing either the full length gene or the specific segment(s) of interest. The template can represent either a single consensus sequence or be a mixture of polymorphic variants of the gene. This DNA template, or scaffold, will preferably include one or more dUTP or rNTP sites in its length. This will provide a means of removing the template prior to carrying out subsequent analytical steps to avoid its acting as a substrate or target in later analysis assays. This removal is accomplished by treating the sample with uracil-DNA glycosylase (UDG) and heating it to cause strand breaks where UDG has generated abasic sites. In the case of rNTP's, the sample can be heated in the presence of a basic buffer (pH ~10) to induce strand breaks where rNTP's are located in the template.

The single stranded DNA template is mixed with the purified RNA, the mixture is denatured and annealed so that the RNA fragments complementary to the DNA template effectively become primers that can be extended along the single stranded DNA templates. DNA polymerase I requires a primer for extension but will efficiently use either a DNA or an RNA primer. Therefore in the presence of DNA polymerase I and dNTP's, the fragmented RNA can be extended along the complementary DNA templates. In order to increase the efficiency of the extension, this reaction can be thermally cycled, allowing overlapping templates and extension products to hybridize and extend until the overall population of fragmented RNA becomes represented as double stranded DNA extended from RNA fragment primers.

Following the generation of this "repaired" RNA, the sample should be treated with UDG or heat-treated in a mildly based solution to fragment the DNA template (scaffold) and prevent it from participating in subsequent analytical reactions.

The product resulting from this enzyme extension can then be used as a template in a standard enzyme profiling assay that includes amplification and detectable signal generation such as fluorescent, chemiluminescent, colorimetric or other common read outs from enzyme based assays. For example, for TaqMan® type assays, this double stranded DNA product is added as the template in a standard assay; and, for array hybridization, this product acts as the cDNA

template for the cRNA labeling reaction typically used to generate single-stranded, labeled RNA for array hybridization.

This method of preparing template has the advantage of recovering information from mRNA fragments too short to effectively act as templates in standard cDNA generation schemes. In addition, this method acts to preserve the specific locations in mRNA sequences targeted by specific analysis assays. For example, TaqMan® assays rely on a single contiguous sequence in a cDNA copy of mRNA to act as a PCR amplification template targeted by a labeled reporter probe. If mRNA strand breaks occur in this sequence, the assay will not detect that template and will underestimate the quantity of that mRNA in the original sample. This target preparation method minimizes that effect from RNA fragmentation.

The extension product formed in the RNA primer extension assay can be controlled by controlling the input quantity of the single stranded DNA template and by doing limited cycling of the extension reaction. This is important in preserving the relative abundance of the mRNA sequences targeted for analysis.

This method has the added advantage of not requiring parallel preparation for each target sequence since it is easily multiplexed. It is also possible to use large pools of random sequence long oligonucleotides or full libraries of cloned sequences to extend the entire population of mRNA sequences in the sample extract for whole expressed genome analysis rather than targeted gene specific analysis.

10. *Amplification of mRNA Species Prior to RT-PCR*

Due to the limited amount and poor quality of mRNA that can be isolated from FPET, a new procedure that could accurately amplify mRNAs of interest would be very useful, particularly for real time quantitation of gene expression (TaqMan®) and especially for quantitatively large number (>50) of genes >50 to 10,000.

Current protocols (e.g. Eberwine, *Biotechniques* 20:584-91 (1996)) are optimized for mRNA amplification from small amount of total or poly A^+ RNA mainly for microarray analysis. The present invention provides a protocol optimized for amplification of small amounts of fragmented total RNA (average size about 60-150 bps), utilizing gene-specific sequences as primers, as illustrated in Figure 4.

The amplification procedure of the invention uses a very large number, typically as many as 100 - 190,000 gene specific primers (GSP's) in one reverse transcription run. Each GSP contains an RNA polymerase promoter, e.g. a T7 DNA-dependent RNA polymerase promoter, at the 5' end for subsequent RNA amplification. GSP's are preferred as primers because of the small size of the RNA. Current protocols utilize dT primers, which would not adequately

represent all reverse transcripts of mRNAs due to the small size of the FPET RNA. GSP's can be designed by optimizing usual parameters, such as length, Tm, etc. For example, GSP's can be designed using the Primer Express® (Applied Biosystems), or Primer 3 (MIT) software program. Typically at least 3 sets per gene are designed, and the ones giving the lowest Ct on FPET RNA (best performers) are selected.

Second strand cDNA synthesis is performed by standard procedures (see Figure 4, Method 1), or by GSP_f primers and Taq pol under PCR conditions (e.g., 95 °C, 10 min (Taq activation) then 60 °C, 45 sec). The advantages of the latter method are that the second gene specific primer, SGF_f adds additional specificity (and potentially more efficient second strand synthesis) and the option of performing several cycles of PCR, if more starting DNA is necessary for RNA amplification by T7 RNA polymerase. RNA amplification is then performed under standard conditions to generate multiple copies of cRNA, which is then used in a standard TaqMan® reaction.

Although this process is illustrated by using T7-based RNA amplification, a person skilled in the art will understand that other RNA polymerase promoters that do not require a primer, such as T3 or Sp6 can also be used, and are within the scope of the invention.

11. *A method of Elongation of Fragmented RNA and Subsequent Amplification*

This method, which combines and modifies the inventions described in sections 9 and 10 above, is illustrated in Figure 5. The procedure begins with elongation of fragmented mRNA. This occurs as described above except that the scaffold DNAs are tagged with the T7 RNA polymerase promoter sequence at their 5' ends, leading to double-stranded DNA extended from RNA fragments. The template sequences need to be removed after *in vitro* transcription. These templates can include dUTP or rNTP nucleotides, enabling enzymatic removal of the templates as described in section 9, or the templates can be removed by DNaseI treatment.

The template DNA can be a population representing different mRNAs of any number. A high sequence complexity source of DNA templates (scaffolds) can be generated by pooling RNA from a variety of cells or tissues. In one embodiment, these RNAs are converted into double stranded DNA and cloned into phagemids. Single stranded DNA can then be rescued by phagemid growth and single stranded DNA isolation from purified phagemids.

This invention is useful because it increases gene expression profile signals two different ways: both by increasing test mRNA polynucleotide sequence length and by *in vitro* transcription amplification. An additional advantage is that it eliminates the need to carry out reverse transcription optimization with gene specific primers tagged with the T7 RNA polymerase promoter sequence, and thus, is comparatively fast and economical.

This invention can be used with a variety of different methods to profile gene expression, e.g., RT-PCR or a variety of DNA array methods. Just as in the previous protocol, this approach is illustrated by using a T7 promoter but the invention is not so limited. A person skilled in the art will appreciate, however, that other RNA polymerase promoters, such as T3 or Sp6 can also be used.

12. *Breast Cancer Gene Set, Assayed Gene Subsequences, and Clinical Application of Gene Expression Data*

An important aspect of the present invention is to use the measured expression of certain genes by breast cancer tissue to match patients to best drugs or drug combinations, and to provide prognostic information. For this purpose it is necessary to correct for (normalize away) both differences in the amount of RNA assayed and variability in the quality of the RNA used. Therefore, the assay measures and incorporates the expression of certain normalizing genes, including well known housekeeping genes, such as GAPDH and Cyp1. Alternatively, normalization can be based on the mean or median signal (Ct) of all of the assayed genes or a large subset thereof (global normalization approach). On a gene-by-gene basis, measured normalized amount of a patient tumor mRNA is compared to the amount found in a breast cancer tissue reference set. The number (N) of breast cancer tissues in this reference set should be sufficiently high to ensure that different reference sets (as a whole) behave essentially the same way. If this condition is met, the identity of the individual breast cancer tissues present in a particular set will have no significant impact on the relative amounts of the genes assayed. Usually, the breast cancer tissue reference set consists of at least about 30, preferably at least about 40 different FPE breast cancer tissue specimens. Unless noted otherwise, normalized expression levels for each mRNA/tested tumor/patient will be expressed as a percentage of the expression level measured in the reference set. More specifically, the reference set of a sufficiently high number (e.g. 40) tumors yields a distribution of normalized levels of each mRNA species. The level measured in a particular tumor sample to be analyzed falls at some percentile within this range, which can be determined by methods well known in the art. Below, unless noted otherwise, reference to expression levels of a gene assume normalized expression relative to the reference set although this is not always explicitly stated.

The breast cancer gene set is shown in Table 1. The gene Accession Numbers, and the SEQ ID NOs for the forward primer, reverse primer and amplicon sequences that can be used for gene amplification, are listed in Table 2. The basis for inclusion of markers, as well as the clinical significance of mRNA level variations with respect to the reference set, is indicated below. Genes are grouped into subsets based on the type of clinical significance indicated by

their expression levels: A. Prediction of patient response to drugs used in breast cancer treatment, or to drugs that are approved for other indications and could be used off-label in the treatment of breast cancer. B. Prognostic for survival or recurrence of cancer.

C. Prediction of Patient Response to Therapeutic Drugs

1. *Molecules that specifically influence cellular sensitivity to drugs*

Table 1 lists 74 genes (shown in italics) that specifically influence cellular sensitivity to potent drugs, which are also listed. Most of the drugs shown are approved and already used to treat breast cancer (e.g., anthracyclines; cyclophosphamide; methotrexate; 5-FU and analogues). Several of the drugs are used to treat breast cancer off-label or are in clinical development phase (e.g., bisphosphonates and anti-VEGF mAb). Several of the drugs have not been widely used to treat breast cancer but are used in other cancers in which the indicated target is expressed (e.g., Celebrex is used to treat familial colon cancer; cisplatin is used to treat ovarian and other cancers.)

Patient response to 5FU is indicated if normalized thymidylate synthase mRNA amount is at or below the 15th percentile, or the sum of expression of thymidylate synthase plus dihydropyrimidine phosphorylase is at or below the 25th percentile, or the sum of expression of these mRNAs plus thymidine phosphorylase is at or below the 20th percentile. Patients with dihydropyrimidine dehydrogenase below 5th percentile are at risk of adverse response to 5FU, or analogs such as Xeloda.

When levels of thymidylate synthase, and dihydropyrimidine dehydrogenase, are within the acceptable range as defined in the preceding paragraph, amplification of c-myc mRNA in the upper 15%, against a background of wild-type p53 [as defined below] predicts a beneficial response to 5FU (see D. Arango *et al.*, *Cancer Res.* 61:4910-4915 (2001)). In the presence of normal levels of thymidylate synthase and dihydropyrimidine dehydrogenase, levels of NFκB and cIAP2 in the upper 10% indicate resistance of breast tumors to the chemotherapeutic drug 5FU.

Patient resistance to anthracyclines is indicated if the normalized mRNA level of topoisomerase IIα is below the 10th percentile, or if the topoisomerase IIβ normalized mRNA level is below the 10th percentile or if the combined normalized topoisomerase IIα and β signals are below the 10th percentile.

Patient sensitivity to methotrexate is compromised if DHFR levels are more than tenfold higher than the average reference set level for this mRNA species, or if reduced folate carrier levels are below 10th percentile.

Patients whose tumors express CYP1B1 in the upper 10%, have reduced likelihood of responding to docetaxol.

The sum of signals for aldehyde dehydrogenase 1A1 and 1A3, when more than tenfold higher than the reference set average, indicates reduced likelihood of response to cyclophosphamide.

Currently, estrogen and progesterone receptor expression as measured by immunohistochemistry is used to select patients for anti-estrogen therapy. We have demonstrated RT-PCR assays for estrogen and progesterone receptor mRNA levels that predict levels of these proteins as determined by a standard clinical diagnostic tests, with high degree of concordance (Figures 6 and 7).

Patients whose tumors express ERα or PR mRNA in the upper 70%, are likely to respond to tamoxifen or other anti-estrogens (thus, operationally, lower levels of ERα than this are to defined ERα-negative). However, when the signal for microsomal epoxide hydrolase is in the upper 10% or when mRNAs for pS2/trefoil factor, GATA3 or human chorionic gonadotropin are at or below average levels found in ERα-negative tumors, anti-estrogen therapy will not be beneficial.

Absence of XIST signal compromises the likelihood of response to taxanes, as does elevation of the GST-π or prolyl endopeptidase [PREP] signal in the upper 10%. Elevation of PLAG1 in the upper 10% decreases sensitivity to taxanes.

Expression of ERCC1 mRNA in the upper 10% indicate significant risk of resistance to cisplatin or analogs.

An RT-PCR assay of Her2 mRNA expression predicts Her2 overexpression as measured by a standard diagnostic test, with high degree of concordance (data not shown). Patients whose tumors express Her2 (normalized to cyp.1) in the upper 10% have increased likelihood of beneficial response to treatment with Herceptin or other ErbB2 antagonists. Measurement of expression of Grb7 mRNA serves as a test for HER2 gene amplification, because the Grb7 gene is closely linked to Her2. When Her2 is expression is high as defined above in this paragraph, similarly elevated Grb7 indicates Her2 gene amplification. Overexpression of IGF1R and or IGF1 or IGF2 decreases likelihood of beneficial response to Herceptin and also to EGFR antagonists.

Patients whose tumors express mutant Ha-Ras, and also express farnesyl pyrophosphate synthetase or geranyl pyrophosphonate synthetase mRNAs at levels above the tenth percentile comprise a group that is especially likely to exhibit a beneficial response to bis-phosphonate drugs.

Cox2 is a key control enzyme in the synthesis of prostaglandins. It is frequently expressed at elevated levels in subsets of various types of carcinomas including carcinoma of the breast. Expression of this gene is controlled at the transcriptional level, so RT-PCR serves a valid indicator of the cellular enzyme activity. Nonclinical research has shown that cox2 promotes tumor angiogenesis, suggesting that this enzyme is a promising drug target in solid tumors. Several Cox2 antagonists are marketed products for use in anti-inflammatory conditions. Treatment of familial adenomatous polyposis patients with the cox2 inhibitor Celebrex significantly decreased the number and size of neoplastic polyps. No cox2 inhibitor has yet been approved for treatment of breast cancer, but generally this class of drugs is safe and could be prescribed off-label in breast cancers in which cox2 is over-expressed. Tumors expressing COX2 at levels in the upper ten percentile have increased chance of beneficial response to Celebrex or other cyclooxygenase 2 inhibitors.

The tyrosine kinases ErbB1 [EGFR], ErbB3 [Her3] and ErbB4 [Her4]; also the ligands TGFalpha, amphiregulin, heparin-binding EGF-like growth factor, and epiregulin; also BRK, a non-receptor kinase. Several drugs in clinical development block the EGF receptor. ErbB2-4, the indicated ligands, and BRK also increase the activity of the EGFR pathway. Breast cancer patients whose tumors express high levels of EGFR or EGFR and abnormally high levels of the other indicated activators of the EGFR pathway are potential candidates for treatment with an EGFR antagonist.

Patients whose tumors express less than 10% of the average level of EGFR mRNA observed in the reference panel are relatively less likely to respond to EGFR antagonists [such as Iressa, or ImClone 225]. In cases in which the EGFR is above this low range, the additional presence of epiregulin, TGFα, amphiregulin, or ErbB3, or BRK, CD9, MMP9, or Lot1 at levels above the 90th percentile predisposes to response to EGFR antagonists. Epiregulin gene expression, in particular, is a good surrogate marker for EGFR activation, and can be used to not only to predict response to EGFR antagonists, but also to monitor response to EGFR antagonists [taking fine needle biopsies to provide tumor tissue during treatment]. Levels of CD82 above the 90th percentile suggest poorer efficacy from EGFR antagonists.

The tyrosine kinases abl, c-kit, PDGFRalpha, PDGFbeta, and ARG; also, the signal transmitting ligands c-kit ligand, PDGFA, B, C and D. The listed tyrosine kinases are all targets of the drug Gleevec™ (imatinib mesylate, Novartis), and the listed ligands stimulate one or more of the listed tyrosine kinases. In the two indications for which Gleevec™ is approved, tyrosine kinase targets (bcr-abl and ckit) are overexpressed and also contain activating mutations. A finding that one of the Gleevec™ target tyrosine kinase targets is expressed in breast cancer tissue

will prompt a second stage of analysis wherein the gene will be sequenced to determine whether it is mutated. That a mutation found is an activating mutation can be proved by methods known in the art, such as, for example, by measuring kinase enzyme activity or by measuring phosphorylation status of the particular kinase, relative to the corresponding wild-type kinase. Breast cancer patients whose tumors express high levels of mRNAs encoding Gleevec™ target tyrosine kinases, specifically, in the upper ten percentile, or mRNAs for Gleevec™ target tyrosine kinases in the average range and mRNAs for their cognate growth stimulating ligands in the upper ten percentile, are particularly good candidates for treatment with Gleevec™.

VEGF is a potent and pathologically important angiogenic factor. (See below under Prognostic Indicators.) When VEGF mRNA levels are in the upper ten percentile, aggressive treatment is warranted. Such levels particularly suggest the value of treatment with anti-angiogenic drugs, including VEGF antagonists, such as anti-VEGF antibodies. Additionally, KDR or CD31 mRNA level in the upper 20 percentile further increases likelihood of benefit from VEGF antagonists.

Farnesyl pyrophosphatase synthetase and geranyl geranyl pyrophosphatase synthetase. These enzymes are targets of commercialized bisphosphonate drugs, which were developed originally for treatment of osteoporosis but recently have begun to prescribe them off-label in breast cancer. Elevated levels of mRNAs encoding these enzymes in breast cancer tissue, above the 90th percentile, suggest use of bisphosphonates as a treatment option.

2. *Multidrug Resistance Factors*

These factors include 10 Genes: gamma glutamyl cysteine synthetase [GCS]; GST-α; GST-π; MDR-1; MRP1-4; breast cancer resistance protein [BCRP]; lung resistance protein [MVP]; SXR; YB-1.

GCS and both GST-α and GST-π regulate glutathione levels, which decrease cellular sensitivity to chemotherapeutic drugs and other toxins by reductive derivatization. Glutathione is a necessary cofactor for multi-drug resistant pumps, MDR-1 and the MRPs. MDR1 and MRPs function to actively transport out of cells several important chemotherapeutic drugs used in breast cancer.

GSTs, MDR-1, and MRP-1 have all been studied extensively to determine possible have prognostic or predictive significance in human cancer. However, a great deal of disagreement exists in the literature with respect to these questions. Recently, new members of the MRP family have been identified: MRP-2, MRP-3, MRP-4, BCRP, and lung resistance protein [major vault protein]. These have substrate specificities that overlap with those of MDR-1 and MRP-1. The incorporation of all of these relevant ABC family members as well as glutathione synthetic

enzymes into the present invention captures the contribution of this family to drug resistance, in a way that single or double analyte assays cannot.

MRP-1, the gene coding for the multidrug resistance protein.

P-glycoprotein, is not regulated primarily at the transcriptional level. However, p-glycoprotein stimulates the transcription of PTP1b. An embodiment of the present invention is the use of the level of the mRNA for the phosphatase PTP1b as a surrogate measure of MRP-1/p-glycoprotein activity.

The gene SXR is also an activator of multidrug resistance, as it stimulates transcription of certain multidrug resistance factors.

The impact of multidrug resistance factors with respect to chemotherapeutic agents used in breast cancer is as follows. Beneficial response to doxorubicin is compromised when the mRNA levels of either MDR1, GSTα, GSTπ, SXR, BCRP YB-1, or LRP/MVP are in the upper four percentile. Beneficial response to methotrexate is inhibited if mRNA levels of any of MRP1, MRP2, MRP3, or MRP4 or gamma-glutamyl cysteine synthetase are in the upper four percentile.

3. *Eukaryotic Translation Initiation Factor 4E [EIF4E]*

EIF4E mRNA levels provides evidence of protein expression and so expands the capability of RT-PCR to indicate variation in gene expression. Thus, one claim of the present invention is the use of EIF4E as an added indicator of gene expression of certain genes [e.g., cyclinD1, mdm2, VEGF, and others]. For example, in two tissue specimens containing the same amount of normalized VEGF mRNA, it is likely that the tissue containing the higher normalized level of EIF4E exhibits the greater level of VEGF gene expression.

The background is as follows. A key point in the regulation of mRNA translation is selection of mRNAs by the EIF4G complex to bind to the 43S ribosomal subunit. The protein EIF4E [the m7G CAP-binding protein] is often limiting because more mRNAs than EIF4E copies exist in cells. Highly structured 5'UTRs or highly GC-rich ones are inefficiently translated, and these often code for genes that carry out functions relevant to cancer [e.g., cyclinD1, mdm2, and VEGF]. EIF4E is itself regulated at the transcriptional/ mRNA level. Thus, expression of EIF4E provides added indication of increased activity of a number of proteins.

It is also noteworthy that overexpression of EIF4E transforms cultured cells, and hence is an oncogene. Overexpression of EIF4E occurs in several different types of carcinomas but is particularly significant in breast cancer. EIF4E is typically expressed at very low levels in normal breast tissue.

D. Prognostic Indicators

1. *DNA Repair Enzymes*

Loss of BRCA1 or BRCA2 activity via mutation represents the critical oncogenic step in the most common type[s] of familial breast cancer. The levels of mRNAs of these important enzymes are abnormal in subsets of sporadic breast cancer as well. Loss of signals from either [to within the lower ten percentile] heightens risk of short survival.

2. *Cell Cycle Regulators*

Cell cycle regulators include 14 genes: c-MYC; c-Src; Cyclin D1; Ha-Ras; mdm2; p14ARF; p21WAF1/CIP; p16INK4a/p14; p23; p27; p53; PI3K; PKC-epsilon; PKC-delta.

The gene for p53 [TP53] is mutated in a large fraction of breast cancers. Frequently p53 levels are elevated when loss of function mutation occurs. When the mutation is dominant-negative, it creates survival value for the cancer cell because growth is promoted and apoptosis is inhibited. Thousands of different p53 mutations have been found in human cancer, and the functional consequences of many of them are not clear. A large body of academic literature addresses the prognostic and predictive significance of mutated p53 and the results are highly conflicting. The present invention provides a functional genomic measure of p53 activity, as follows. The activated wild type p53 molecule triggers transcription of the cell cycle inhibitor p21. Thus, the ratio of p53 to p21 should be low when p53 is wild-type and activated. When p53 is detectable and the ratio of p53 to p21 is elevated in tumors relative to normal breast, it signifies nonfunctional or dominant negative p53. The cancer literature provides evidence for this as born out by poor prognosis.

Mdm2 is an important p53 regulator. Activated wildtype p53 stimulates transcription of mdm2. The mdm2 protein binds p53 and promotes its proteolytic destruction. Thus, abnormally low levels of mdm2 in the presence of normal or higher levels of p53 indicate that p53 is mutated and inactivated.

One aspect of the present invention is the use of ratios of mRNAs levels p53:p21 and p53:mdm2 to provide a picture of p53 status. Evidence for dominant negative mutation of p53 (as indicated by high p53:p21 and/or high p53:mdm2 mRNA ratios—specifically in the upper ten percentile) presages higher risk of recurrence in breast cancer and therefore weights toward a decision to use chemotherapy in node negative post surgery breast cancer.

Another important cell cycle regulator is p27, which in the activated form blocks cell cycle progression at the level of cdk4. The protein is regulated primarily via phosphorylation/dephosphorylation, rather than at the transcriptional level. However, levels of

p27 mRNAs do vary. Therefore a level of p27 mRNA in the upper ten percentile indicates reduced risk of recurrence of breast cancer post surgery.

Cyclin D1 is a principle positive regulator of entry into S phase of the cell cycle. The gene for cyclin D1 is amplified in about 20% of breast cancer patients, and therefore promotes tumor promotes tumor growth in those cases. One aspect of the present invention is use of cyclin D1 mRNA levels for diagnostic purposes in breast cancer. A level of cyclin D1 mRNA in the upper ten percentile suggests high risk of recurrence in breast cancer following surgery and suggests particular benefit of adjuvant chemotherapy.

3. *Other tumor suppressors and related proteins*

These include APC and E-cadherin. It has long been known that the tumor suppressor APC is lost in about 50% of colon cancers, with concomitant transcriptional upregulation of E-cadherin, an important cell adhesion molecule and growth suppressor. Recently, it has been found that the APC gene silenced in 15-40 % of breast cancers. Likewise, the E-cadherin gene is silenced [via CpG island methylation] in about 30% of breast cancers. An abnormally low level of APC and/or E-cadherin mRNA in the lower 5 percentile suggests high risk of recurrence in breast cancer following surgery and heightened risk of shortened survival.

4. *Regulators of Apoptosis*

These include BCl/BAX family members BCl2, Bcl-xl, Bak, Bax and related factors, NFκ-B and related factors, and also p53BP1/ASPP1 and p53BP2/ASPP2.

Bax and Bak are pro-apoptotic and BCl2 and Bcl-xl are anti-apoptotic. Therefore, the ratios of these factors influence the resistance or sensitivity of a cell to toxic (pro-apoptotic) drugs. In breast cancer, unlike other cancers, elevated level of BCl2 (in the upper ten percentile) correlates with good outcome. This reflects the fact that BCl2 has growth inhibitory activity as well as anti-apoptotic activity, and in breast cancer the significance of the former activity outweighs the significance of the latter. The impact of BCl2 is in turn dependent on the status of the growth stimulating transcription factor c-MYC. The gene for c-MYC is amplified in about 20% of breast cancers. When c-MYC message levels are abnormally elevated relative to BCl2 (such that this ratio is in the upper ten percentile), then elevated level of BCl2 mRNA is no longer a positive indicator.

NFκ-B is another important anti-apoptotic factor. Originally, recognized as a pro-inflammatory transcription factor, it is now clear that it prevents programmed cell death in response to several extracellular toxic factors [such as tumor necrosis factor]. The activity of this transcription factor is regulated principally via phosphorylation/dephosphorylation events. However, levels of NFκ-B nevertheless do vary from cell to cell, and elevated levels should

correlate with increased resistance to apoptosis. Importantly for present purposes, NFκ-B, exerts its anti-apoptotic activity largely through its stimulation of transcription of mRNAs encoding certain members of the IAP [inhibitor of apoptosis] family of proteins, specifically cIAP1, cIAP2, XIAP, and Survivin. Thus, abnormally elevated levels of mRNAs for these IAPs and for NFκ-B any in the upper 5 percentile] signify activation of the NFκ-B anti-apoptotic pathway. This suggests high risk of recurrence in breast cancer following chemotherapy and therefore poor prognosis. One embodiment of the present invention is the inclusion in the gene set of the above apoptotic regulators, and the above-outlined use of combinations and ratios of the levels of their mRNAs for prognosis in breast cancer.

The proteins p53BP1 and 2 bind to p53 and promote transcriptional activation of pro-apoptotic genes. The levels of p53BP1 and 2 are suppressed in a significant fraction of breast cancers, correlating with poor prognosis. When either is expressed in the lower tenth percentile poor prognosis is indicated.

5. *Factors that control cell invasion and angiogenesis*

These include uPA, PAI1, cathepsinsB, G and L, scatter factor [HGF], c-met, KDR, VEGF, and CD31. The plasminogen activator uPA and its serpin regulator PAI1 promote breakdown of extracellular matrices and tumor cell invasion. Abnormally elevated levels of both mRNAs in malignant breast tumors (in the upper twenty percentile) signify an increased risk of shortened survival, increased recurrence in breast cancer patients post surgery, and increased importance of receiving adjuvant chemotherapy. On the other hand, node negative patients whose tumors do not express elevated levels of these mRNA species are less likely to have recurrence of this cancer and could more seriously consider whether the benefits of standard chemotherapy justifies the associated toxicity.

Cathepsins B or L, when expressed in the upper ten percentile, predict poor disease-free and overall survival. In particular, cathepsin L predicts short survival in node positive patients.

Scatter factor and its cognate receptor c-met promote cell motility and invasion, cell growth, and angiogenesis. In breast cancer elevated levels of mRNAs encoding these factors should prompt aggressive treatment with chemotherapeutic drugs, when expression of either, or the combination, is above the 90^{th} percentile.

VEGF is a central positive regulator of angiogenesis, and elevated levels in solid tumors predict short survival [note many references showing that elevated level of VEGF predicts short survival]. Inhibitors of VEGF therefore slow the growth of solid tumors in animals and humans. VEGF activity is controlled at the level of transcription. VEGF mRNA levels in the upper ten percentile indicate significantly worse than average prognosis. Other markers of vascularization,

CD31 [PECAM], and KDR indicate high vessel density in tumors and that the tumor will be particularly malignant and aggressive, and hence that an aggressive therapeutic strategy is warranted.

6. *Markers for Immune and Inflammatory Cells and Processes*

These markers include the genes for Immunoglobulin light chain λ, CD18, CD3, CD68, Fas [CD95], and Fas Ligand.

Several lines of evidence suggest that the mechanisms of action of certain drugs used in breast cancer entail activation of the host immune/inflammatory response (For example, Herceptin®). One aspect of the present invention is the inclusion in the gene set of markers for inflammatory and immune cells, and markers that predict tumor resistance to immune surveillance. Immunoglobulin light chain lambda is a marker for immunoglobulin producing cells. CD18 is a marker for all white cells. CD3 is a marker for T-cells. CD68 is a marker for macrophages.

CD95 and Fas ligand are a receptor: ligand pair that mediate one of two major pathways by which cytotoxic T cells and NK cells kill targeted cells. Decreased expression of CD95 and increased expression of Fas Ligand indicates poor prognosis in breast cancer. Both CD95 and Fas Ligand are transmembrane proteins, and need to be membrane anchored to trigger cell death. Certain tumor cells produce a truncated soluble variant of CD95, created as a result of alternative splicing of the CD95 mRNA. This blocks NK cell and cytotoxic T cell Fas Ligand-mediated killing of the tumors cells. Presence of soluble CD95 correlates with poor survival in breast cancer. The gene set includes both soluble and full-length variants of CD95.

7. *Cell proliferation markers*

The gene set includes the cell proliferation markers Ki67/MiB1, PCNA, Pin1, and thymidine kinase. High levels of expression of proliferation markers associate with high histologic grade, and short survival. High levels of thymidine kinase in the upper ten percentile suggest in creased risk of short survival. Pin1 is a prolyl isomerase that stimulates cell growth, in part through the transcriptional activation of the cyclin D1 gene, and levels in the upper ten percentile contribute to a negative prognostic profile.

8. *Other growth factors and receptors*

This gene set includes IGF1, IGF2, IGFBP3, IGF1R, FGF2, FGFR1, CSF-1R/fms, CSF-1, IL6 and IL8. All of these proteins are expressed in breast cancer. Most stimulate tumor growth. However, expression of the growth factor FGF2 correlates with good outcome. Some have anti-apoptotic activity, prominently IGF1. Activation of the IGF1 axis via elevated IGF1, IGF1R, or

IGFBP3 (as indicated by the sum of these signals in the upper ten percentile) inhibits tumor cell death and strongly contributes to a poor prognostic profile.

9. *Gene expression markers that define subclasses of breast cancer*

These include: GRO1 oncogene alpha, Grb7, cytokeratins 5 and 17, retinal binding protein 4, hepatocyte nuclear factor 3, integrin alpha 7, and lipoprotein lipase. These markers subset breast cancer into different cell types that are phenotypically different at the level of gene expression. Tumors expressing signals for Bcl2, hepatocyte nuclear factor 3, LIV1 and ER above the mean have the best prognosis for disease free and overall survival following surgical removal of the cancer. Another category of breast cancer tumor type, characterized by elevated expression of lipoprotein lipase, retinol binding protein 4, and integrin α7, carry intermediate prognosis. Tumors expressing either elevated levels of cytokeratins 5, and 17, GRO oncogene at levels four-fold or greater above the mean, or ErbB2 and Grb7 at levels ten-fold or more above the mean, have worst prognosis.

Although throughout the present description, including the Examples below, various aspects of the invention are explained with reference to gene expression studies, the invention can be performed in a similar manner, and similar results can be reached by applying proteomics techniques that are well known in the art. The proteome is the totality of the proteins present in a sample (e.g. tissue, organism, or cell culture) at a certain point of time. Proteomics includes, among other things, study of the global changes of protein expression in a sample (also referred to as "expression proteomics"). Proteomics typically includes the following steps: (1) separation of individual proteins in a sample by 2-D gel electrophoresis (2-D PAGE); (2) identification of the individual proteins recovered from the gel, e.g. my mass spectrometry and/or N-terminal sequencing, and (3) analysis of the data using bioinformatics. Proteomics methods are valuable supplements to other methods of gene expression profiling, and can be used, alone or in combination with other methods of the present invention, to detect the products of the gene markers of the present invention.

Further details of the invention will be described in the following non-limiting Examples.

Example 1

Isolation of RNA from formalin-fixed, paraffin-embedded (FPET) tissue specimens

A. Protocols

I. EPICENTRE® Xylene Protocol

RNA Isolation

(1) Cut 1-6 sections (each 10 μm thick) of paraffin-embedded tissue per sample using a clean microtome blade and place into a 1.5 ml eppendorf tube.

(2) To extract paraffin, add 1 ml of xylene and invert the tubes for 10 minutes by rocking on a nutator.

(3) Pellet the sections by centrifugation for 10 minutes at 14,000 x g in an eppendorf microcentrifuge.

(4) Remove the xylene, leaving some in the bottom to avoid dislodging the pellet.

(5) Repeat steps 2-4.

(6) Add 1 ml of 100% ethanol and invert for 3 minutes by rocking on the nutator.

(7) Pellet the debris by centrifugation for 10 minutes at 14,000 x g in an eppendorf microcentrifuge.

(8) Remove the ethanol, leaving some at the bottom to avoid the pellet.

(9) Repeat steps 6-8 twice.

(10) Remove all of the remaining ethanol.

(11) For each sample, add 2 µl of 50 µg/µl Proteinase K to 300 µl of Tissue and Cell Lysis Solution.

(12) Add 300 µl of Tissue and Cell Lysis Solution containing the Proteinase K to each sample and mix thoroughly.

(13) Incubate at 65 °C for 90 minutes (vortex mixing every 5 minutes). Visually monitor the remaining tissue fragment. If still visible after 30 minutes, add an additional 2 µl of 50 µg/µl Proteinase K and continue incubating at 65 °C until fragment dissolves.

(14) Place the samples on ice for 3-5 minutes and proceed with protein removal and total nucleic acid precipitation.

<u>Protein Removal and Precipitation of Total Nucleic Acid</u>

(1) Add 150 µl of MPC Protein Precipitation Reagent to each lysed sample and vortex vigorously for 10 seconds.

(2) Pellet the debris by centrifugation for 10 minutes at 14,000 x g in an eppendorf microcentrifuge.

(3) Transfer the supernatant into clean eppendorf tubes and discard the pellet.

(4) Add 500 µl of isopropanol to the recovered supernatant and thoroughly mix by rocking on the nutator for 3 minutes.

(5) Pellet the RNA/DNA by centrifugation at 4 °C for 10 minutes at 14,000 x g in an eppendorf microcentrifuge.

(6) Remove all of the isopropanol with a pipet, being careful not to dislodge the pellet.

Removal of Contaminating DNA from RNA Preparations

(1) Prepare 200 µl of DNase I solution for each sample by adding 5 µl of RNase-Free DNase I (1 U/µl) to 195 µl of 1X DNase Buffer.

(2) Completely resuspend the pelleted RNA in 200 µl of DNase I solution by vortexing.

(3) Incubate the samples at 37 °C for 60 minutes.

(4) Add 200 µl of 2X T and C Lysis Solution to each sample and vortex for 5 seconds.

(5) Add 200 µl of MPC Protein Precipitation Reagent, mix by vortexing for 10 seconds and place on ice for 3-5 minutes.

(6) Pellet the debris by centrifugation for 10 minutes at 14,000 x g in an eppendorf microcentrifuge.

(7) Transfer the supernatant containing the RNA to clean eppendorf tubes and discard the pellet. (Be careful to avoid transferring the pellet.)

(8) Add 500 µl of isopropanol to each supernatant and rock samples on the nutator for 3 minutes.

(9) Pellet the RNA by centrifugation at 4 °C for 10 minutes at 14,000 x g in an eppendorf microcentrifuge.

(10) Remove the isopropanol, leaving some at the bottom to avoid dislodging the pellet.

(11) Rinse twice with 1 ml of 75% ethanol. Centrifuge briefly if the RNA pellet is dislodged.

(12) Remove ethanol carefully.

(13) Set under fume hood for about 3 minutes to remove residual ethanol.

(14) Resuspend the RNA in 30 µl of TE Buffer and store at -30 °C.

II. Hot Wax/Urea Protocol of the Invention

RNA Isolation

(1) Cut 3 sections (each 10 µm thick) of paraffin-embedded tissue using a clean microtome blade and place into a 1.5 ml eppendorf tube.

(2) Add 300 µl of lysis buffer (10 mM Tris 7.5, 0.5% sodium lauroyl sarcosine, 0.1 mM EDTA pH 7.5, 4M Urea) containing 330 µg/ml Proteinase K (added freshly from a 50 µg/µl stock solution) and vortex briefly.

(3) Incubate at 65 °C for 90 minutes (vortex mixing every 5 minutes). Visually monitor the tissue fragment. If still visible after 30 minutes, add an additional 2 µl of 50 µg/µl Proteinase K and continue incubating at 65 °C until fragment dissolves.

(4) Centrifuge for 5 minutes at 14,000 x g and transfer upper aqueous phase to new tube, being careful not to disrupt the paraffin seal.

(5) Place the samples on ice for 3-5 minutes and proceed with protein removal and total nucleic acid precipitation.

Protein Removal and Precipitation of Total Nucleic Acid

(1) Add 150 µl of 7.5M NH_4OAc to each lysed sample and vortex vigorously for 10 seconds.

(2) Pellet the debris by centrifugation for 10 minutes at 14,000 x g in an eppendorf microcentrifuge.

(3) Transfer the supernatant into clean eppendorf tubes and discard the pellet.

(4) Add 500 µl of isopropanol to the recovered supernatant and thoroughly mix by rocking on the nutator for 3 minutes.

(5) Pellet the RNA/DNA by centrifugation at 4 °C for 10 minutes at 14,000 x g in an eppendorf microcentrifuge.

(6) Remove all of the isopropanol with a pipet, being careful not to dislodge the pellet.

Removal of Contaminating DNA from RNA Preparations

(1) Add 45 µl of 1X DNase I buffer (10 mM Tris-Cl, pH 7,5, 2.5 mM $MgCl_2$, 0.1 mM $CaCl_2$) and 5 µl of RNase-Free DNase I (2U/µl, Ambion) to each sample.

(2) Incubate the samples at 37 °C for 60 minutes.

Inactivate the DNaseI by heating at 70 °C for 5 minutes.

B. Results

Experimental evidence demonstrates that the hot RNA extraction protocol of the invention does not compromise RNA yield. Using 19 FPE breast cancer specimens, extracting RNA from three adjacent sections in the same specimens, RNA yields were measured via capillary electrophoresis with fluorescence detection (Agilent Bioanalyzer). Average RNA yields in nanograms and standard deviations with the invented and commercial methods, respectively, were: 139+/-21 versus 141+/-34.

Also, it was found that the urea-containing lysis buffer of the present invention can be substituted for the EPICENTRE® T&C lysis buffer, and the 7.5 M NH_4OAc reagent used for protein precipitation in accordance with the present invention can be substituted for the

EPICENTRE® MPC protein precipitation solution with neither significant compromise of RNA yield nor TaqMan® efficiency.

Example 2

Amplification of mRNA Species Prior to RT-PCR

The method described in section 10 above was used with RNA isolated from fixed, paraffin-embedded breast cancer tissue. TaqMan® analyses were performed with first strand cDNA generated with the T7-GSP primer (unamplified (T7-GSPr)), T7 amplified RNA (amplified (T7-GSPr)). RNA was amplified according to step 2 of Figure 4. As a control, TaqMan® was also performed with cDNA generated with an unmodified GSPr (amplified (GSPr)). An equivalent amount of initial template (1 ng/well) was used in each TaqMan® reaction.

The results are shown in Figure 8. *In vitro* transcription increased RT-PCR signal intensity by more than 10 fold, and for certain genes by more than 100 fold relative to controls in which the RT-PCR primers were the same primers used in method 2 for the generation of double-stranded DNA for *in vitro* transcription (GSP-$T7_r$ and GSP_f). Also shown in Figure 8 are RT-PCR data generated when standard optimized RT-PCR primers (i.e., lacking T7 tails) were used. As shown, compared to this control, the new method yielded substantial increases in RT-PCR signal (from 4 to 64 fold in this experiment).

The new method requires that each T7-GSP sequence be optimized so that the increase in the RT-PCR signal is the same for each gene, relative to the standard optimized RT-PCR (with non-T7 tailed primers).

Example 3

A Study of Gene Expression in Premalignant and Malignant Breast Tumors

A gene expression study was designed and conducted with the primary goal to molecularly characterize gene expression in paraffin-embedded, fixed tissue samples of invasive breast ductal carcinoma, and to explore the correlation between such molecular profiles and disease-free survival. A further objective of the study was to compare the molecular profiles in tissue samples of invasive breast cancer with the molecular profiles obtained in ductal carcinoma *in situ*. The study was further designed to obtain data on the molecular profiles in lobular carcinoma *in situ* and in paraffin-embedded, fixed tissue samples of invasive lobular carcinoma.

Molecular assays were performed on paraffin-embedded, formalin-fixed primary breast tumor tissues obtained from 202 individual patients diagnosed with breast cancer. All patients underwent surgery with diagnosis of invasive ductal carcinoma of the breast, pure ductal carcinoma in situ (DCIS), lobular carcinoma of the breast, or pure lobular carcinoma in situ

(LCIS). Patients were included in the study only if histopathologic assessment, performed as described in the Materials and Methods section, indicated adequate amounts of tumor tissue and homogeneous pathology.

The individuals participating in the study were divided into the following groups:

Group 1: Pure ductal carcinoma in situ (DCIS); n=18

Group 2: Invasive ductal carcinoma n=130

Group 3: Pure lobular carcinoma in situ (LCIS); n=7

Group 4: Invasive lobular carcinoma n=16

Materials and Methods

Each representative tumor block was characterized by standard histopathology for diagnosis, semi-quantitative assessment of amount of tumor, and tumor grade. A total of 6 sections (10 microns in thickness each) were prepared and placed in two Costar Brand Microcentrifuge Tubes (Polypropylene, 1.7 mL tubes, clear; 3 sections in each tube). If the tumor constituted less than 30% of the total specimen area, the sample may have been crudely dissected by the pathologist, using gross microdissection, putting the tumor tissue directly into the Costar tube.

If more than one tumor block was obtained as part of the surgical procedure, all tumor blocks were subjected to the same characterization, as described above, and the block most representative of the pathology was used for analysis.

Gene Expression Analysis

mRNA was extracted and purified from fixed, paraffin-embedded tissue samples, and prepared for gene expression analysis as described in chapters 7-11 above.

Molecular assays of quantitative gene expression were performed by RT-PCR, using the ABI PRISM 7900™ Sequence Detection System™ (Perkin-Elmer-Applied Biosystems, Foster City, CA, USA). ABI PRISM 7900™ consists of a thermocycler, laser, charge-coupled device (CCD), camera and computer. The system amplifies samples in a 384-well format on a thermocycler. During amplification, laser-induced fluorescent signal is collected in real-time through fiber optics cables for all 384 wells, and detected at the CCD. The system includes software for running the instrument and for analyzing the data.

Analysis and Results

Tumor tissue was analyzed for 185 cancer-related genes and 7 reference genes. The threshold cycle (CT) values for each patient were normalized based on the median of all genes

for that particular patient. Clinical outcome data were available for all patients from a review of registry data and selected patient charts.

Outcomes were classified as:

0 died due to breast cancer or to unknown cause or alive with breast cancer recurrence;

1 alive without breast cancer recurrence or died due to a cause other than breast cancer

Analysis was performed by:

1. Analysis of the relationship between normalized gene expression and the binary outcomes of 0 or 1.

2. Analysis of the relationship between normalized gene expression and the time to outcome (0 or 1 as defined above) where patients who were alive without breast cancer recurrence or who died due to a cause other than breast cancer were censored. This approach was used to evaluate the prognostic impact of individual genes and also sets of multiple genes.

Analysis of 147 patients with invasive breast carcinoma by binary approach

In the first (binary) approach, analysis was performed on all 146 patients with invasive breast carcinoma. A t test was performed on the group of patients classified as 0 or 1 and the p-values for the differences between the groups for each gene were calculated.

The following Table 4 lists the 45 genes for which the p-value for the differences between the groups was <0.05.

Table 4

Gene/ SEQ ID NO:	Mean CT Alive	Mean CT Deceased	t-value	Degrees of freedom	p
FOXM1	33.66	32.52	3.92	144	0.0001
PRAME	35.45	33.84	3.71	144	0.0003
Bcl2	28.52	29.32	-3.53	144	0.0006
STK15	30.82	30.10	3.49	144	0.0006
CEGP1	29.12	30.86	-3.39	144	0.0009
Ki-67	30.57	29.62	3.34	144	0.0011
GSTM1	30.62	31.63	-3.27	144	0.0014
CA9	34.96	33.54	3.18	144	0.0018
PR	29.56	31.22	-3.16	144	0.0019
BBC3	31.54	32.10	-3.10	144	0.0023
NME1	27.31	26.68	3.04	144	0.0028
SURV	31.64	30.68	2.92	144	0.0041
GATA3	26.06	26.99	-2.91	144	0.0042
TFRC	28.96	28.48	2.87	144	0.0047
YB-1	26.72	26.41	2.79	144	0.0060
DPYD	28.51	28.84	-2.67	144	0.0084
GSTM3	28.21	29.03	-2.63	144	0.0095
RPS6KB1	31.18	30.61	2.61	144	0.0099
Src	27.97	27.69	2.59	144	0.0105
Chk1	32.63	31.99	2.57	144	0.0113
ID1	28.73	29.13	-2.48	144	0.0141
EstR1	24.22	25.40	-2.44	144	0.0160
p27	27.15	27.51	-2.41	144	0.0174
CCNB1	31.63	30.87	2.40	144	0.0176
XIAP	30.27	30.51	-2.40	144	0.0178
Chk2	31.48	31.11	2.39	144	0.0179
CDC25B	29.75	29.39	2.37	144	0.0193
IGF1R	28.85	29.44	-2.34	144	0.0209

AK055699	33.23	34.11	-2.28	144	0.0242
PI3KC2A	31.07	31.42	-2.25	144	0.0257
TGFB3	28.42	28.85	-2.25	144	0.0258
BAGI1	28.40	28.75	-2.24	144	0.0269
CYP3A4	35.70	35.32	2.17	144	0.0317
EpCAM	28.73	28.34	2.16	144	0.0321
VEGFC	32.28	31.82	2.16	144	0.0326
pS2	28.96	30.60	-2.14	144	0.0341
hENT1	27.19	26.91	2.12	144	0.0357
WISP1	31.20	31.64	-2.10	144	0.0377
HNF3A	27.89	28.64	-2.09	144	0.0384
NFKBp65	33.22	33.80	-2.08	144	0.0396
BRCA2	33.06	32.62	2.08	144	0.0397
EGFR	30.68	30.13	2.06	144	0.0414
TK1	32.27	31.72	2.02	144	0.0453
VDR	30.08	29.73	1.99	144	0.0488

In the foregoing Table 4, lower (negative) t-values indicate higher expression (or lower CTs), associated with better outcomes, and, inversely, higher (positive) t-values indicate higher expression (lower CTs) associated with worse outcomes. Thus, for example, elevated expression of the FOXM1 gene (t-value = 3.92, CT mean alive> CT mean deceased) indicates a reduced likelihood of disease free survival. Similarly, elevated expression of the CEGP1 gene (t-value = -3.39; CT mean alive< CT mean deceased) indicates an increased likelihood of disease free survival.

Based on the data set forth in Table 4, the overexpression of any of the following genes in breast cancer indicates a reduced likelihood of survival without cancer recurrence following surgery: FOXM1; PRAME; SKT15, Ki-67; CA9; NME1; SURV; TFRC; YB-1; RPS6KB1; Src; Chk1; CCNB1; Chk2; CDC25B; CYP3A4; EpCAM; VEGFC; hENT1; BRCA2; EGFR; TK1; VDR.

Based on the data set forth in Table 4, the overexpression of any of the following genes in breast cancer indicates a better prognosis for survival without cancer recurrence following surgery: Blc12; CEGP1; GSTM1; PR; BBC3; GATA3; DPYD; GSTM3; ID1; EstR1; p27; XIAP; IGF1R; AK055699; P13KC2A; TGFB3; BAGI1; pS2; WISP1; HNF3A; NFKBp65.

Analysis of 108 ER positive patient by binary approach

108 patients with normalized CT for estrogen receptor (ER) < 25.2 (i.e., ER positive patients) were subjected to separate analysis. A t test was performed on the groups of patients classified as 0 or 1 and the p-values for the differences between the groups for each gene were calculated. The following Table 5 lists the 12 genes where the p-value for the differences between the groups was <0.05.

Table 5

Gene/ SEQ ID NO:	Mean CT Alive	Mean CT Deceased	t-value	Degrees of freedom	p
PRAME	35.54	33.88	3.03	106	0.0031
Bcl2	28.24	28.87	-2.70	106	0.0082
FOXM1	33.82	32.85	2.66	106	0.089
DIABLO	30.33	30.71	-2.47	106	0.0153
EPHX1	28.62	28.03	2.44	106	0.0163
HIF1A	29.37	28.88	2.40	106	0.0180
VEGFC	32.39	31.69	2.39	106	0.0187
Ki-67	30.73	29.82	2.38	106	0.0191
IGF1R	28.60	29.18	-2.37	106	0.0194
VDR	30.14	29.60	2.17	106	0.0322
NME1	27.34	26.80	2.03	106	0.0452
GSTM3	28.08	28.92	-2.00	106	0.0485

For each gene, a classification algorithm was utilized to identify the best threshold value (CT) for using each gene alone in predicting clinical outcome.

Based on the data set forth in Table 5, overexpression of the following genes in ER-positive cancer is indicative of a reduced likelihood of survival without cancer recurrence following surgery: PRAME; FOXM1; EPHX1; HIF1A; VEGFC; Ki-67; VDR; NME1. Some of these genes (PRAME; FOXM1; VEGFC; Ki-67; VDR; and NME1) were also identified as indicators of poor prognosis in the previous analysis, not limited to ER-positive breast cancer. The overexpression of the remaining genes (EPHX1 and HIF1A) appears to be negative indicator of disease free survival in ER-positive breast cancer only. Based on the data set forth in Table 5, overexpression of the following genes in ER-positive cancer is indicative of a better

prognosis for survival without cancer recurrence following surgery: Bcl-2; DIABLO; IGF1R; GSTM3. Of the latter genes, Bcl-2; IGFR1; and GSTM3 have also been identified as indicators of good prognosis in the previous analysis, not limited to ER-positive breast cancer. The overexpression of DIABLO appears to be positive indicator of disease free survival in ER-positive breast cancer only.

Analysis of multiple genes and indicators of outcome

Two approaches were taken in order to determine whether using multiple genes would provide better discrimination between outcomes.

First, a discrimination analysis was performed using a forward stepwise approach. Models were generated that classified outcome with greater discrimination than was obtained with any single gene alone.

According to a second approach (time-to-event approach), for each gene a Cox Proportional Hazards model (see, e.g. Cox, D. R., and Oakes, D. (1984), *Analysis of Survival Data*, Chapman and Hall, London, New York) was defined with time to recurrence or death as the dependent variable, and the expression level of the gene as the independent variable. The genes that have a p-value < 0.05 in the Cox model were identified. For each gene, the Cox model provides the relative risk (RR) of recurrence or death for a unit change in the expression of the gene. One can choose to partition the patients into subgroups at any threshold value of the measured expression (on the CT scale), where all patients with expression values above the threshold have higher risk, and all patients with expression values below the threshold have lower risk, or vice versa, depending on whether the gene is an indicator of good ($RR>1.01$) or poor ($RR<1.01$) prognosis. Thus, any threshold value will define subgroups of patients with respectively increased or decreased risk. The results are summarized in the following Tables 6 and 7.

Table 6

Cox Model Results for 146 Patients with Invasive Breast Cancer

Gene	Relative Risk (RR)	SE Relative Risk	p value
FOXM1	0.58	0.15	0.0002
STK15	0.51	0.20	0.0006
PRAME	0.78	0.07	0.0007
Bcl2	1.66	0.15	0.0009
CEGP1	1.25	0.07	0.0014
GSTM1	1.40	0.11	0.0014
Ki67	0.62	0.15	0.0016
PR	1.23	0.07	0.0017
Contig51037	0.81	0.07	0.0022
NME1	0.64	0.15	0.0023
YB-1	0.39	0.32	0.0033
TFRC	0.53	0.21	0.0035
BBC3	1.72	0.19	0.0036
GATA3	1.32	0.10	0.0039
CA9	0.81	0.07	0.0049
SURV	0.69	0.13	0.0049
DPYD	2.58	0.34	0.0052
RPS6KB1	0.60	0.18	0.0055
GSTM3	1.36	0.12	0.0078
Src.2	0.39	0.36	0.0094
TGFB3	1.61	0.19	0.0109
CDC25B	0.54	0.25	0.0122
XIAP	3.20	0.47	0.0126
CCNB1	0.68	0.16	0.0151
IGF1R	1.42	0.15	0.0153
Chk1	0.68	0.16	0.0155
ID1	1.80	0.25	0.0164
p27	1.69	0.22	0.0168
Chk2	0.52	0.27	0.0175

EstR1	1.17	0.07	0.0196
HNF3A	1.21	0.08	0.206
pS2	1.12	0.05	0.0230
BAGl1	1.88	0.29	0.0266
AK055699	1.24	0.10	0.0276
pENT1	0.51	0.31	0.0293
EpCAM	0.62	0.22	0.0310
WISP1	1.39	0.16	0.0338
VEGFC	0.62	0.23	0.0364
TK1	0.73	0.15	0.0382
NFKBp65	1.32	0.14	0.0384
BRCA2	0.66	0.20	0.0404
CYP3A4	0.60	0.25	0.0417
EGFR	0.72	0.16	0.0436

Table 7

Cox Model Results for 108 Patients wih ER+ Invasive Breast Cancer

Gene	Relative Risk (RR)	SE Relative Risk	p-value
PRAME	0.75	0.10	0.0045
Contig51037	0.75	0.11	0.0060
Blc2	2.11	0.28	0.0075
HIF1A	0.42	0.34	0.0117
IGF1R	1.92	0.26	0.0117
FOXM1	0.54	0.24	0.0119
EPHX1	0.43	0.33	0.0120
Ki67	0.60	0.21	0.0160
CDC25B	0.41	0.38	0.0200
VEGFC	0.45	0.37	0.0288
CTSB	0.32	0.53	0.0328
DIABLO	2.91	0.50	0.0328
p27	1.83	0.28	0.0341
CDH1	0.57	0.27	0.0352
IGFBP3	0.45	0.40	0.0499

The binary and time-to-event analyses, with few exceptions, identified the same genes as prognostic markers. For example, comparison of Tables 4 and 6 shows that, with the exception of a single gene, the two analyses generated the same list of top 15 markers (as defined by the smallest p values). Furthermore, when both analyses identified the same gene, they were concordant with respect to the direction (positive or negative sign) of the correlation with survival/recurrence. Overall, these results strengthen the conclusion that the identified markers have significant prognostic value.

For Cox models comprising more than two genes (multivariate models), stepwise entry of each individual gene into the model is performed, where the first gene entered is pre-selected from among those genes having significant univariate p-values, and the gene selected for entry into the model at each subsequent step is the gene that best improves the fit of the model to the data. This analysis can be performed with any total number of genes. In the analysis the results of which are shown below, stepwise entry was performed for up to 10 genes.

Multivariate analysis is performed using the following equation:

RR=exp[coef(geneA) x Ct(geneA) + coef(geneB) x Ct(geneB) + coef(geneC) x Ct(geneC) + ...].

In this equation, coefficiencts for genes that are predictors of beneficial outcome are positive numbers and coefficients for genes that are predictors of unfavorable outcome are negative numbers. The "Ct" values in the equation are ΔCts, i.e. reflect the difference between the average normalized Ct value for a population and the normalized Ct measured for the patient in question. The convention used in the present analysis has been that ΔCts below and above the population average have positive signs and negative signs, respectively (reflecting greater or lesser mRNA abundance). The relative risk (RR) calculated by solving this equation will indicate if the patient has an enhanced or reduced chance of long-term survival without cancer recurrence.

Multivariate gene analysis of 147 patients with invasive breast carcinoma

(a) A multivariate stepwise analysis, using the Cox Proportional Hazards Model, was performed on the gene expression data obtained for all 147 patients with invasive breast carcinoma. Genes CEGP1, FOXM1, STK15 and PRAME were excluded from this analysis. The following ten-gene sets have been identified by this analysis as having particularly strong predictive value of patient survival without cancer recurrence following surgical removal of primary tumor.

1. Bcl2, cyclinG1, NFKBp65, NME1, EPHX1, TOP2B, DR5, TERC, Src, DIABLO;

2. Ki67, XIAP, hENT1, TS, CD9, p27, cyclinG1, pS2, NFKBp65, CYP3A4;

3. GSTM1, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, NFKBp65, ErbB3;

4. PR, NME1, XIAP, upa, cyclinG1, Contig51037, TERC, EPHX1, ALDH1A3, CTSL;

5. CA9, NME1, TERC, cyclinG1, EPHX1, DPYD, Src, TOP2B, NFKBp65, VEGFC;

6. TFRC, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, ErbB3, NFKBp65.

(b) A multivariate stepwise analysis, using the Cox Proportional Hazards Model, was performed on the gene expression data obtained for all 147 patients with invasive breast carcinoma, using an interrogation set including a reduced number of genes. The following ten-gene sets have been identified by this analysis as having particularly strong predictive value of patient survival without cancer recurrence following surgical removal of primary tumor.

1. Bcl2, PRAME, cyclinG1, FOXM1, NFKBp65, TS, XIAP, Ki67, CYP3A4, p27;

2. FOXM1, cyclinG1, XIAP, Contig51037, PRAME, TS, Ki67, PDGFRa, p27, NFKBp65;

3. PRAME, FOXM1, cyclinG1, XIAP, Contig51037, TS, Ki6, PDGFRa, p27, NFKBp65;

4. Ki67, XIAP, PRAME, hENT1, contig51037, TS, CD9, p27, ErbB3, cyclinG1;

5. STK15, XIAP, PRAME, PLAUR, p27, CTSL, CD18, PREP, p53, RPS6KB1;

6. GSTM1, XIAP, PRAME, p27, Contig51037, ErbB3, GSTp, EREG, ID1, PLAUR;

7. PR, PRAME, NME1, XIAP, PLAUR, cyclinG1, Contig51037, TERC, EPHX1, DR5;

8. CA9, FOXM1, cyclinG1, XIAP, TS, Ki67, NFKBp65, CYP3A4, GSTM3, p27;

9. TFRC, XIAP, PRAME, p27, Contig51037, ErbB3, DPYD, TERC, NME1, VEGFC;

10. CEGP1, PRAME, hENT1, XIAP, Contig51037, ErbB3, DPYD, NFKBp65, ID1, TS.

Multivariate analysis of patients with ER positive invasive breast carcinoma

A multivariate stepwise analysis, using the Cox Proportional Hazards Model, was performed on the gene expression data obtained for patients with ER positive invasive breast carcinoma. The following ten-gene sets have been identified by this analysis as having particularly strong predictive value of patient survival without cancer recurrence following surgical removal of primary tumor.

1. PRAME, p27, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;
2. Contig51037, EPHX1, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;
3. Bcl2, hENT1, FOXM1, Contig51037, cyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;
4. HIF1A, PRAME, p27, IGFBP2, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;
5. IGF1R, PRAME, EPHX1, Contig51037, cyclinG1, Bcl2, NME1, PTEN, TBP, TIMP2;
6. FOXM1, Contig51037, VEGFC, TBP, HIF1A, DPYD, RAD51C, DCR3, cyclinG1, BAG1;
7. EPHX1, Contig51037, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;
8. Ki67, VEGFC, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;
9. CDC25B, Contig51037, hENT1, Bcl2, HLAG, TERC, NME1, upa, ID1, CYP;
10. VEGFC, Ki67, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;
11. CTSB, PRAME, p27, IGFBP2, EPHX1, CTSL, BAD, DR5, DCR3, XIAP;
12. DIABLO, Ki67, hENT1, TIMP2, ID1, p27, KRT19, IGFBP2, TS, PDGFB;
13. p27, PRAME, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;
14. CDH1; PRAME, VEGFC; HIF1A; DPYD, TIMP2, CYP3A4, EstR1, RBP4, p27;
15. IGFBP3, PRAME, p27, Bcl2, XIAP, EstR1, Ki67, TS, Src, VEGF;
16. GSTM3, PRAME, p27, IGFBP3, XIAP, FGF2, hENT1, PTEN, EstR1, APC;
17. hENT1, Bcl2, FOXM1, Contig51037, CyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;
18. STK15, VEGFC, PRAME, p27, GCLC, hENT1, ID1, TIMP2, EstR1, MCP1;

19. NME1, PRAM, p27, IGFBP3, XIAP, PTEN, hENT1, Bcl2, CYP3A4, HLAG;

20. VDR, Bcl2, p27, hENT1, p53, PI3KC2A, EIF4E, TFRC, MCM3, ID1;

21. EIF4E, Contig51037, EPHX1, cyclinG1, Bcl2, DR5, TBP, PTEN, NME1, HER2;

22. CCNB1, PRAME, VEGFC, HIF1A, hENT1, GCLC, TIMP2, ID1, p27, upa;

23. ID1, PRAME, DIABLO, hENT1, p27, PDGFRa, NME1, BIN1, BRCA1, TP;

24. FBXO5, PRAME, IGFBP3, p27, GSTM3, hENT1, XIAP, FGF2, TS, PTEN;

25. GUS, HIA1A, VEGFC, GSTM3, DPYD, hENT1, FBXO5, CA9, CYP, KRT18;

26. Bclx, Bcl2, hENT1, Contig51037, HLAG, CD9, ID1, BRCA1, BIN1, HBEGF.

It is noteworthy that many of the foregoing gene sets include genes that alone did not have sufficient predictive value to qualify as prognostic markers under the standards discussed above, but in combination with other genes, their presence provides valuable information about the likelihood of long-term patient survival without cancer recurrence

All references cited throughout the disclosure are hereby expressly incorporated by reference.

While the present invention has been described with reference to what are considered to be the specific embodiments, it is to be understood that the invention is not limited to such embodiments. To the contrary, the invention is intended to cover various modifications and equivalents included within the spirit and scope of the appended claims. For example, while the disclosure focuses on the identification of various breast cancer associated genes and gene sets, and on the diagnosis and treatment of breast cancer, similar genes, gene sets and methods concerning other types of cancer are specifically within the scope herein.

WHAT IS CLAIMED IS:

1. A method for predicting clinical outcome for a patient diagnosed with cancer, comprising

determining the expression level of one or more genes, or their expression products, selected from the group consisting of p53BP2, cathepsin B, cathepsin L, Ki67/MiB1, and thymidine kinase in a cancer tissue obtained from the patient, normalized against a control gene or genes, and compared to the amount found in a reference cancer tissue set,

wherein a poor outcome is predicted if:

(a) the expression level of p53BP2 is in the lower 10^{th} percentile; or

(b) the expression level of either cathepsin B or cathepsin L is in the upper 10^{th} percentile; or

(c) the expression level of any either Ki67/MiB1 or thymidine kinase is in the upper 10^{th} percentile.

2. The method of claim 1 wherein poor clinical outcome is measured in terms of shortened survival or increased risk of cancer recurrence.

3. The method of claim 2 wherein poor clinical outcome is measured in terms of shortened survival or increased risk of cancer recurrence following surgical removal of the cancer.

4. The method of claim 1 wherein the cancer is selected from the group consisting of breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer, pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, and brain cancer.

5. The method of claim 4 wherein the cancer is breast cancer.

6. The method of claim 5 wherein the expression level of p53BP2 is determined.

7. The method of claim 5 wherein the expression levels of cathepsin B and cathepsin L are determined.

8. The method of claim 5 wherein the expression level of cathepsin L is determined.

9. The method of claim 5 wherein the expression levels of Ki67/MiB1 and thymidine kinase are determined.

10. The method of claim 5 wherein the expression level of Ki67/MiB1 is determined.

11. The method of claim 5 wherein the expression level of thymidine kinase is determined.

12. The method of claim 1 wherein the expression level of more than one gene, or gene product, is determined.

13. The method of claim 1 wherein the expression level of more than two genes is determined.

14. The method of claim 13 further comprising the step of subjecting the expression data to multivariate analysis using the Cox Proportional Hazards model.

15. The method of claim 1 wherein the expression level is determined using RNA obtained from a formalin-fixed, paraffin-embedded tissue sample.

16. The method of claim 1 wherein the expression level is determined by reverse phase polymerase chain reaction (RT-PCR).

17. The method of claim 16 wherein said RNA is fragmented.

18. A method of predicting the likelihood of the recurrence of cancer following treatment in a cancer patient, comprising determining the expression level of p27, or its expression product, in a cancer tissue obtained from said patient, normalized against a control gene or genes, and compared to the amount found in a reference cancer tissue set, wherein an expression level in the upper 10th percentile indicates decreased risk of recurrence following treatment.

19. The method of claim 18 wherein the cancer is selected from the group consisting of breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer,

pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, and brain cancer.

20. The method of claim 19 wherein the cancer is breast cancer.

21. The method of claim 20 wherein the expression level is determined following surgical removal of cancer.

22. The method of claim 20 wherein the expression level is determined using RNA obtained from a formalin-fixed, paraffin-embedded tissue sample.

23. The method of claim 22 wherein said RNA is fragmented.

24. The method of claim 22 wherein the expression level is determined by reverse phase polymerase chain reaction (RT-PCR).

25. A method for classifying cancer comprising, determining the expression level of two or more genes selected from the group consisting of Bcl2, hepatocyte nuclear factor 3, ER, ErbB2 and Grb7, or their expression products, in a cancer tissue, normalized against a control gene or genes, and compared to the amount found in a reference cancer tissue set, wherein (i) tumors expressing at least one of Bcl2, hepatocyte nuclear factor 3, and ER, or their expression products, above the mean expression level in the reference tissue set are classified as having a good prognosis for disease free and overall patient survival following treatment; and (ii) tumors expressing elevated levels of ErbB2 and Grb7, or their expression products, at levels ten-fold or more above the mean expression level in the reference tissue set are classified as having poor prognosis of disease free and overall patient survival following treatment.

26. The method of claim 26 wherein the cancer is selected from the group consisting of breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer, pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, and brain cancer.

27. The method of claim 26 wherein the cancer is breast cancer.

28. The method of claim 26 wherein the expression level is determined following surgical removal of cancer.

29. The method of claim 26 wherein the expression level is determined using RNA obtained from a formalin-fixed, paraffin-embedded tissue sample.

30. The method of claim 29 wherein said RNA is fragmented.

31. The method of claim 29 wherein the expression level is determined by reverse phase polymerase chain reaction (RT-PCR).

32. A method of predicting the likelihood of long-term survival of a breast cancer patient without the recurrence of breast cancer, following surgical removal of the primary tumor, comprising determining the expression level of one or more prognostic RNA transcripts or their product in a breast cancer tissue sample obtained from said patient, normalized against the expression level of all RNA transcripts or their products in said breast cancer tissue sample, or of a reference set of RNA transcripts or their products, wherein the prognostic transcript is the transcript of one or more genes selected from the group consisting of: FOXM1, PRAME, Bcl2, STK15, CEGP1, Ki-67, GSTM1, CA9, PR, BBC3, NME1, SURV, GATA3, TFRC, YB-1, DPYD, GSTM3, RPS6KB1, Src, Chk1, ID1, EstR1, p27, CCNB1, XIAP, Chk2, CDC25B, IGF1R, AK055699, P13KC2A, TGFB3, BAGI1, CYP3A4, EpCAM, VEGFC, pS2, hENT1, WISP1, HNF3A, NFKBp65, BRCA2, EGFR, TK1, VDR, Contig51037, pENT1, EPHX1, IF1A, DIABLO, CDH1, HIF1α, IGFBP3, CTSB, and Her2, wherein overexpression of one or more of FOXM1, PRAME, STK15, Ki-67, CA9, NME1, SURV, TFRC, YB-1, RPS6KB1, Src, Chk1, CCNB1, Chk2, CDC25B, CYP3A4, EpCAM, VEGFC, hENT1, BRCA2, EGFR, TK1, VDR, EPHX1, IF1A, Contig51037, CDH1, HIF1α, IGFBP3, CTSB, Her2, and pENT1 indicates a decreased likelihood of long-term survival without breast cancer recurrence, and the overexpression of one or more of Bcl2, CEGP1, GSTM1, PR, BBC3, GATA3, DPYD, GSTM3, ID1, EstR1, p27, XIAP, IGF1R, AK055699, P13KC2A, TGFB3, BAGI1, pS2, WISP1, HNF3A, NFKBp65, and DIABLO indicates an increased likelihood of long-term survival without breast cancer recurrence.

33. The method of claim 32 comprising determining the expression level of at least two of said prognostic transcipts or their expression products.

34. The method of claim 32 wherein the breast cancer is invasive breast carcinoma, comprising determination of the expression levels of the transcripts of the following genes, or their expression products: FOXM1, PRAME, Bcl2, STK15, CEGP1, Ki-67, GSTM1, PR, BBC3, NME1, SURV, GATA3, TFRC, YB-1, DPYD, CA9, Contig51037, RPS6K1 and Her2.

35. The method of claim 32 wherein said breast cancer is characterized by overexpression of the estrogen receptor (ER).

36. The method of claim 35 comprising determination of the expression levels of the transcripts of at least two of the following genes, or their expression products: PRAME, Bcl2, FOXM1, DIABLO, EPHX1, HIF1A, VEGFC, Ki-67, IGF1R, VDR, NME1, GSTM3, Contig51037, CDC25B, CTSB, p27, CDH1, and IGFBP3.

37. The method of claim 32 wherein the expression level of one or more prognostic RNA transcripts is determined.

38. The method of claim 37 wherein said RNA is isolated from a fixed, wax-embedded breast cancer tissue specimen of said patient.

39. An array comprising polynucleotides hybridizing to the following genes: FOXM1, PRAME, Bcl2, STK15, CEGP1, Ki-67, GSTM1, PR, BBC3, NME1, SURV, GATA3, TFRC, YB-1, DPYD, CA9, Contig51037, RPS6K1 and Her2, immobilized on a solid surface.

40. The array of claim 39 comprising polynucleotides hybridizing to the following genes: FOXM1, PRAME, Bcl2, STK15, CEGP1, Ki-67, GSTM1, CA9, PR, BBC3, NME1, SURV, GATA3, TFRC, YB-1, DPYD, GSTM3, RPS6KB1, Src, Chk1, ID1, EstR1, p27, CCNB1, XIAP, Chk2, CDC25B, IGF1R, AK055699, P13KC2A, TGFB3, BAGI1, CYP3A4, EpCAM, VEGFC, pS2, hENT1, WISP1, HNF3A, NFKBp65, BRCA2, EGFR, TK1, VDR, Contig51037, pENT1, EPHX1, IF1A, CDH1, HIF1α, IGFBP3, CTSB, Her2 and DIABLO, immobilized on a solid surface.

41. A method of predicting the likelihood of long-term survival of a patient diagnosed with invasive breast cancer, without the recurrence of breast cancer, following surgical removal of the primary tumor, comprising the steps of:

(1) determining the expression levels of the RNA transcripts or the expression products of genes of a gene set selected from the group consisting of

(a) Bcl2, cyclinG1, NFKBp65, NME1, EPHX1, TOP2B, DR5, TERC, Src, DIABLO;

(b) Ki67, XIAP, hENT1, TS, CD9, p27, cyclinG1, pS2, NFKBp65, CYP3A4;

(c) GSTM1, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, NFKBp65, ErbB3;

(d) PR, NME1, XIAP, upa, cyclinG1, Contig51037, TERC, EPHX1, ALDH1A3, CTSL;

(e) CA9, NME1, TERC, cyclinG1, EPHX1, DPYD, Src, TOP2B, NFKBp65, VEGFC;

(f) TFRC, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, ErbB3, NFKBp65;

(g) Bcl2, PRAME, cyclinG1, FOXM1, NFKBp65, TS, XIAP, Ki67, CYP3A4, p27;

(h) FOXM1, cyclinG1, XIAP, Contig51037, PRAME, TS, Ki67, PDGFRa, p27, NFKBp65;

(i) PRAME, FOXM1, cyclinG1, XIAP, Contig51037, TS, Ki6, PDGFRa, p27, NFKBp65;

(j) Ki67, XIAP, PRAME, hENT1, contig51037, TS, CD9, p27, ErbB3, cyclinG1;

(k) STK15, XIAP, PRAME, PLAUR, p27, CTSL, CD18, PREP, p53, RPS6KB1;

(l) GSTM1, XIAP, PRAME, p27, Contig51037, ErbB3, GSTp, EREG, ID1, PLAUR;

(m) PR, PRAME, NME1, XIAP, PLAUR, cyclinG1, Contig51037, TERC, EPHX1, DR5;

(n) CA9, FOXM1, cyclinG1, XIAP, TS, Ki67, NFKBp65, CYP3A4, GSTM3, p27;

(o) TFRC, XIAP, PRAME, p27, Contig51037, ErbB3, DPYD, TERC, NME1, VEGFC; and

(p) CEGP1, PRAME, hENT1, XIAP, Contig51037, ErbB3, DPYD, NFKBp65, ID1, TS

in a breast cancer tissue sample obtained from said patient, normalized against the expression levels of all RNA transcripts or their products in said breast cancer tissue sample, or of a reference set of RNA transcripts or their products;

(2) subjecting the data obtained in step (a) to statistical analysis; and

(3) determining whether the likelihood of said long-term survival has increased or decreased.

42. A method of predicting the likelihood of long-term survival of a patient diagnosed with estrogen receptor (ER)-positive invasive breast cancer, without the recurrence of breast cancer, following surgical removal of the primary tumor, comprising the steps of:

(1) determining the expression levels of the RNA transcripts or the expression products of genes of a gene set selected from the group consisting of

(a) PRAME, p27, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(b) Contig51037, EPHX1, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;

(c) Bcl2, hENT1, FOXM1, Contig51037, cyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;

(d) HIF1A, PRAME, p27, IGFBP2, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(e) IGF1R, PRAME, EPHX1, Contig51037, cyclinG1, Bcl2, NME1, PTEN, TBP, TIMP2;

(f) FOXM1, Contig51037, VEGFC, TBP, HIF1A, DPYD, RAD51C, DCR3, cyclinG1, BAG1;

(g) EPHX1, Contig51037, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;

(h) Ki67, VEGFC, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;

(i) CDC25B, Contig51037, hENT1, Bcl2, HLAG, TERC, NME1, upa, ID1, CYP;

(j) VEGFC, Ki67, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;

(k) CTSB, PRAME, p27, IGFBP2, EPHX1, CTSL, BAD, DR5, DCR3, XIAP;

(l) DIABLO, Ki67, hENT1, TIMP2, ID1, p27, KRT19, IGFBP2, TS, PDGFB;

(m) p27, PRAME, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(n) CDH1; PRAME, VEGFC; HIF1A; DPYD, TIMP2, CYP3A4, EstR1, RBP4, p27;

(o) IGFBP3, PRAME, p27, Bcl2, XIAP, EstR1, Ki67, TS, Src, VEGF;

(p) GSTM3, PRAME, p27, IGFBP3, XIAP, FGF2, hENT1, PTEN, EstR1, APC;

(q) hENT1, Bcl2, FOXM1, Contig51037, CyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;

(r) STK15, VEGFC, PRAME, p27, GCLC, hENT1, ID1, TIMP2, EstR1, MCP1;

(s) NME1, PRAM, p27, IGFBP3, XIAP, PTEN, hENT1, Bcl2, CYP3A4, HLAG;

(t) VDR, Bcl2, p27, hENT1, p53, PI3KC2A, EIF4E, TFRC, MCM3, ID1;

(u) EIF4E, Contig51037, EPHX1, cyclinG1, Bcl2, DR5, TBP, PTEN, NME1, HER2;

(v) CCNB1, PRAME, VEGFC, HIF1A, hENT1, GCLC, TIMP2, ID1, p27, upa;

(w) ID1, PRAME, DIABLO, hENT1, p27, PDGFRa, NME1, BIN1, BRCA1, TP;

(x). FBXO5, PRAME, IGFBP3, p27, GSTM3, hENT1, XIAP, FGF2, TS, PTEN;

(y) GUS, HIA1A, VEGFC, GSTM3, DPYD, hENT1, EBXO5, CA9, CYP, KRT18; and

(z) Bclx, Bcl2, hENT1, Contig51037, HLAG, CD9, ID1, BRCA1, BIN1, HBEGF;

(2) subjecting the data obtained in step (1) to statistical analysis; and

(3) determining whether the likelihood of said long-term survival has increased or decreased.

43. The method of claim 41 or claim 42 wherein said statistical analysis is performed by using the Cox Proportional Hazards model.

44. An array comprising polynucleotides hybridizing to a gene set selected from the group consisting of

(a) Bcl2, cyclinG1, NFKBp65, NME1, EPHX1, TOP2B, DR5, TERC, Src, DIABLO;

(b) Ki67, XIAP, hENT1, TS, CD9, p27, cyclinG1, pS2, NFKBp65, CYP3A4;

(c) GSTM1, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, NFKBp65, ErbB3;

(d) PR, NME1, XIAP, upa, cyclinG1, Contig51037, TERC, EPHX1, ALDH1A3, CTSL;

(e) CA9, NME1, TERC, cyclinG1, EPHX1, DPYD, Src, TOP2B, NFKBp65, VEGFC;

(f) TFRC, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, ErbB3, NFKBp65;

(g) Bcl2, PRAME, cyclinG1, FOXM1, NFKBp65, TS, XIAP, Ki67, CYP3A4, p27;

(h) FOXM1, cyclinG1, XIAP, Contig51037, PRAME, TS, Ki67, PDGFRa, p27, NFKBp65;

(i) PRAME, FOXM1, cyclinG1, XIAP, Contig51037, TS, Ki6, PDGFRa, p27, NFKBp65;

(j) Ki67, XIAP, PRAME, hENT1, contig51037, TS, CD9, p27, ErbB3, cyclinG1;

(k) STK15, XIAP, PRAME, PLAUR, p27, CTSL, CD18, PREP, p53, RPS6KB1;

(l) GSTM1, XIAP, PRAME, p27, Contig51037, ErbB3, GSTp, EREG, ID1, PLAUR;

(m) PR, PRAME, NME1, XIAP, PLAUR, cyclinG1, Contig51037, TERC, EPHX1, DR5;

(n) CA9, FOXM1, cyclinG1, XIAP, TS, Ki67, NFKBp65, CYP3A4, GSTM3, p27;

(o) TFRC, XIAP, PRAME, p27, Contig51037, ErbB3, DPYD, TERC, NME1, VEGFC; and

(p) CEGP1, PRAME, hENT1, XIAP, Contig51037, ErbB3, DPYD, NFKBp65, ID1, TS,

immobilized on a solid surface.

45. An array comprising polynucleotides hybridizing to a gene set selected from the group consisting of

(a) PRAME, p27, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(b) Contig51037, EPHX1, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;

(c) Bcl2, hENT1, FOXM1, Contig51037, cyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;

(d) HIF1A, PRAME, p27, IGFBP2, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(e) IGF1R, PRAME, EPHX1, Contig51037, cyclinG1, Bcl2, NME1, PTEN, TBP, TIMP2;

(f) FOXM1, Contig51037, VEGFC, TBP, HIF1A, DPYD, RAD51C, DCR3, cyclinG1, BAG1;

(g) EPHX1, Contig51037, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;

(h) Ki67, VEGFC, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;

(i) CDC25B, Contig51037, hENT1, Bcl2, HLAG, TERC, NME1, upa, ID1, CYP;

(j) VEGFC, Ki67, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;

(k) CTSB, PRAME, p27, IGFBP2, EPHX1, CTSL, BAD, DR5, DCR3, XIAP;

(l) DIABLO, Ki67, hENT1, TIMP2, ID1, p27, KRT19, IGFBP2, TS, PDGFB;

(m) p27, PRAME, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(n) CDH1; PRAME, VEGFC; HIF1A; DPYD, TIMP2, CYP3A4, EstR1, RBP4, p27;

(o) IGFBP3, PRAME, p27, Bcl2, XIAP, EstR1, Ki67, TS, Src, VEGF;

(p) GSTM3, PRAME, p27, IGFBP3, XIAP, FGF2, hENT1, PTEN, EstR1, APC;

(q) hENT1, Bcl2, FOXM1, Contig51037, CyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;

(r) STK15, VEGFC, PRAME, p27, GCLC, hENT1, ID1, TIMP2, EstR1, MCP1;

(s) NME1, PRAM, p27, IGFBP3, XIAP, PTEN, hENT1, Bcl2, CYP3A4, HLAG;

(t) VDR, Bcl2, p27, hENT1, p53, PI3KC2A, EIF4E, TFRC, MCM3, ID1;

(u) EIF4E, Contig51037, EPHX1, cyclinG1, Bcl2, DR5, TBP, PTEN, NME1, HER2;

(v) CCNB1, PRAME, VEGFC, HIF1A, hENT1, GCLC, TIMP2, ID1, p27, upa;

(w) ID1, PRAME, DIABLO, hENT1, p27, PDGFRa, NME1, BIN1, BRCA1, TP;

(x) FBXO5, PRAME, IGFBP3, p27, GSTM3, hENT1, XIAP, FGF2, TS, PTEN;

(y) GUS, HIA1A, VEGFC, GSTM3, DPYD, hENT1, FBXO5, CA9, CYP, KRT18; and

(z) Bclx, Bcl2, hENT1, Contig51037, HLAG, CD9, ID1, BRCA1, BIN1, HBEGF,

immobilized on a solid surface.

Overall FPET/RT-PCR Flow Chart

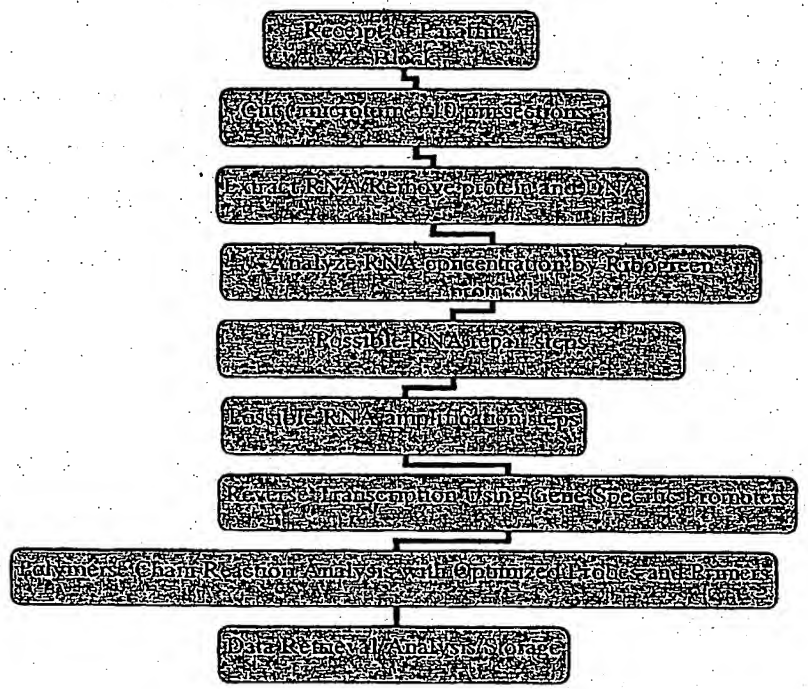


FIGURE 1

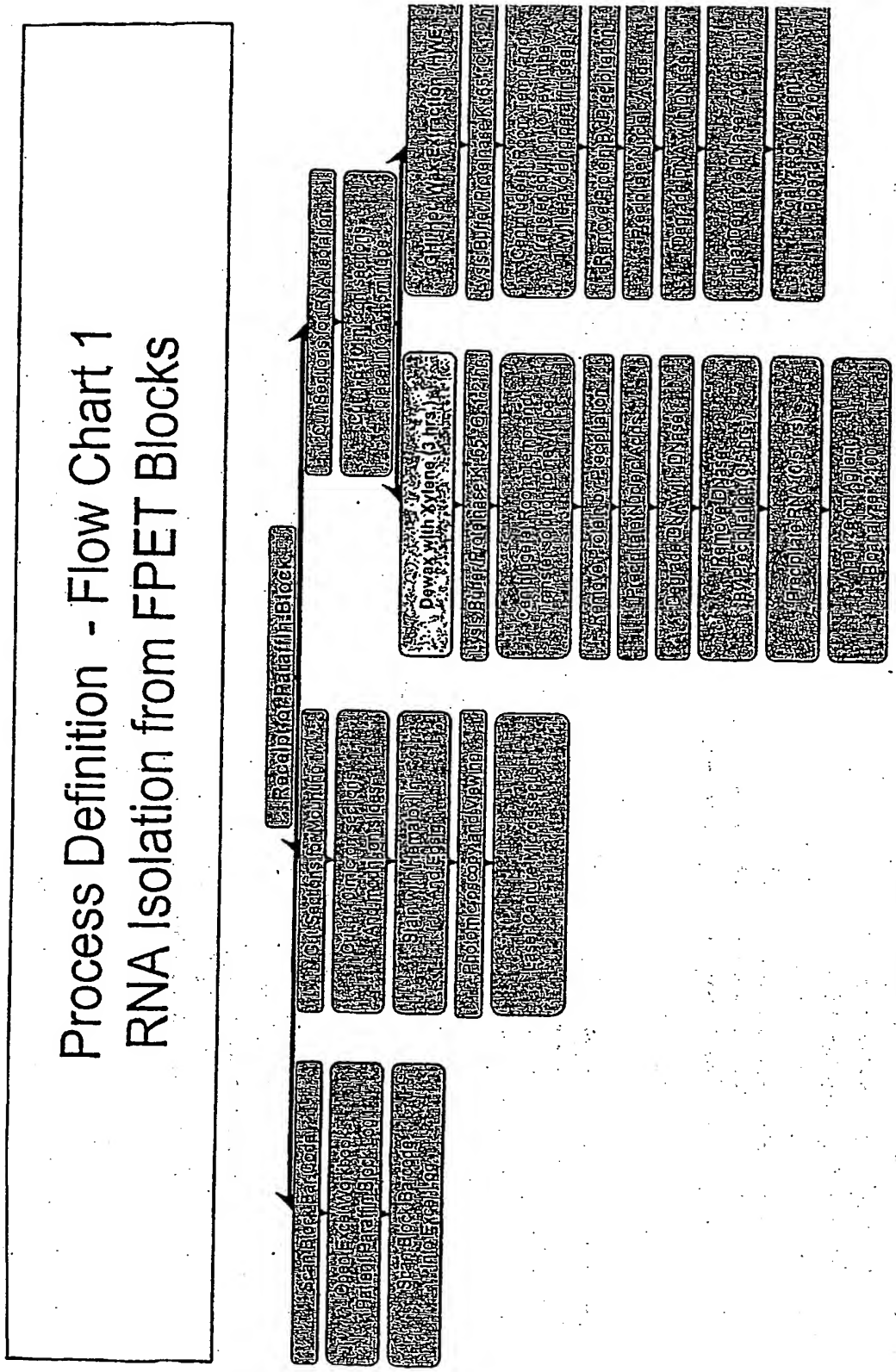

Process Definition - Flow Chart 1
RNA Isolation from FPET Blocks
Dewax with Xylene (3 hrs.)
Estimated
total time for
20 samples
12 hrs
8 hrs

FIGURE 2

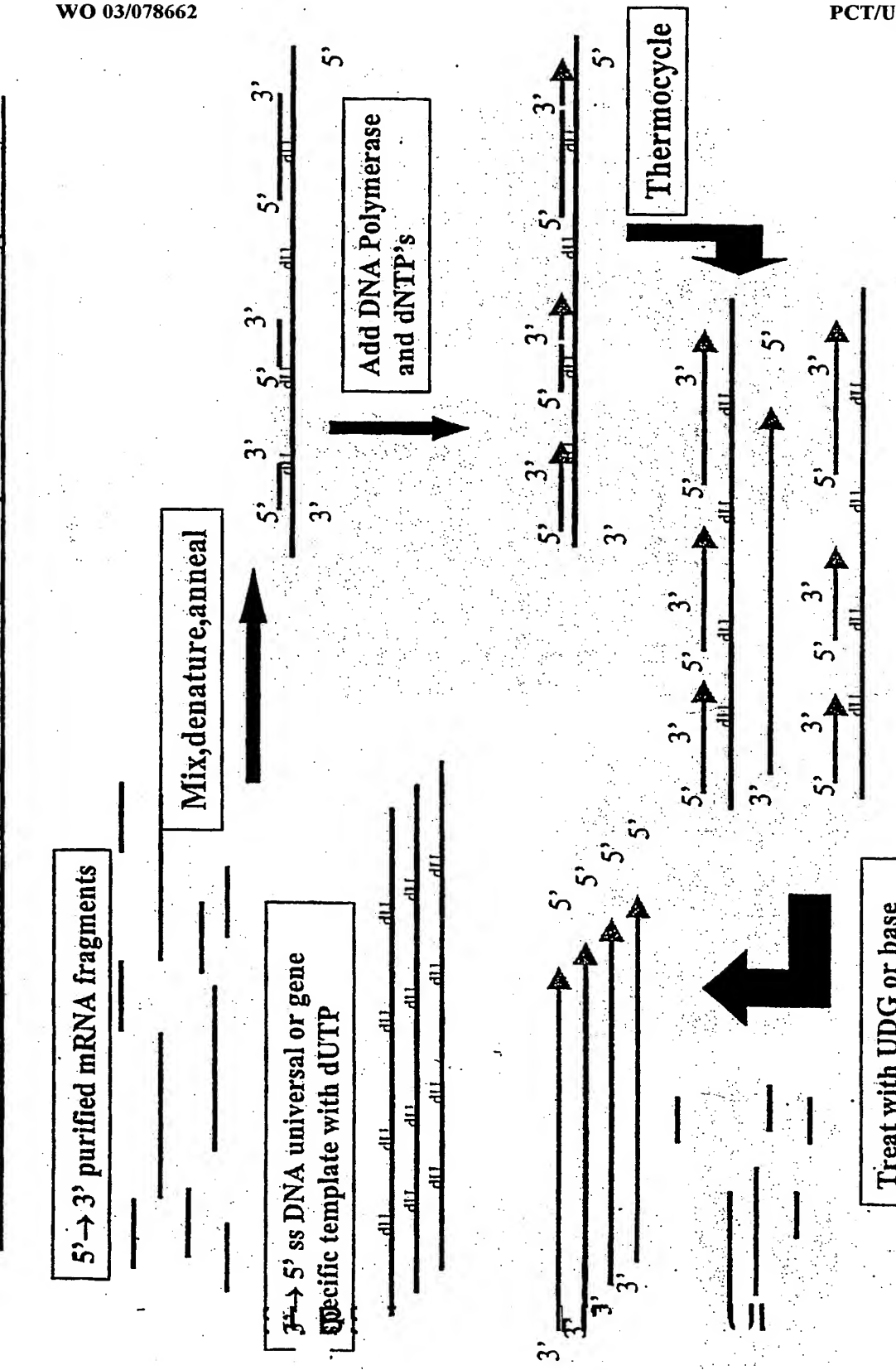
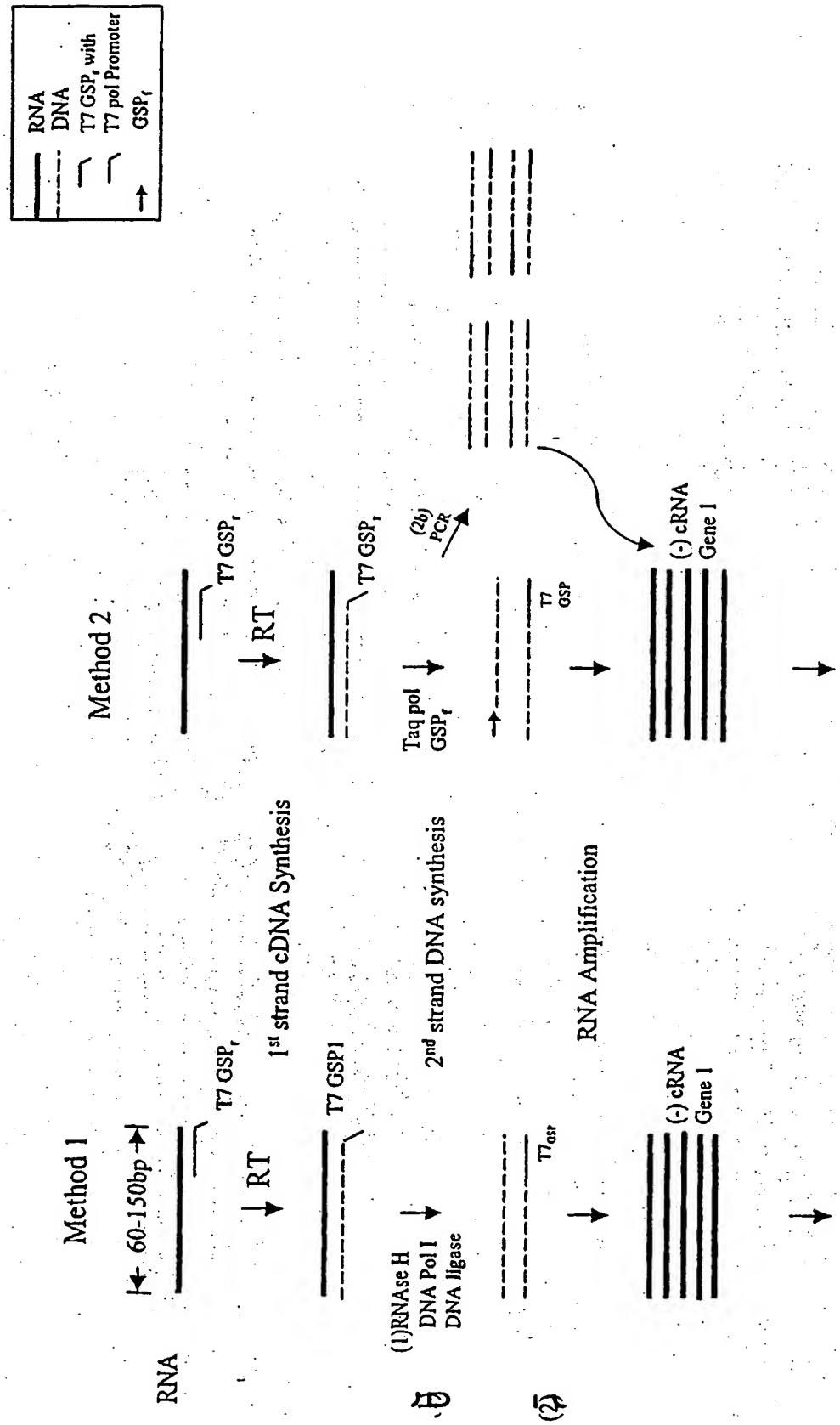

Scheme for Preparing Fragmented mRNA for Expression Profiling Analysis
5'→ 3' purified mRNA fragments
Mix,denature,anneal
3'→ 5' ss DNA universal or gene specific template with dUTP
Add DNA Polymerase and dNTP's
Thermocycle
Treat with UDG or base To fragment DNA scaffold

FIGURE 3



RT-PCR (one-step or two-step)

FIGURE 4

Alternative Scheme for Preparing Fragmented mRNA for Expression Profiling Analysis

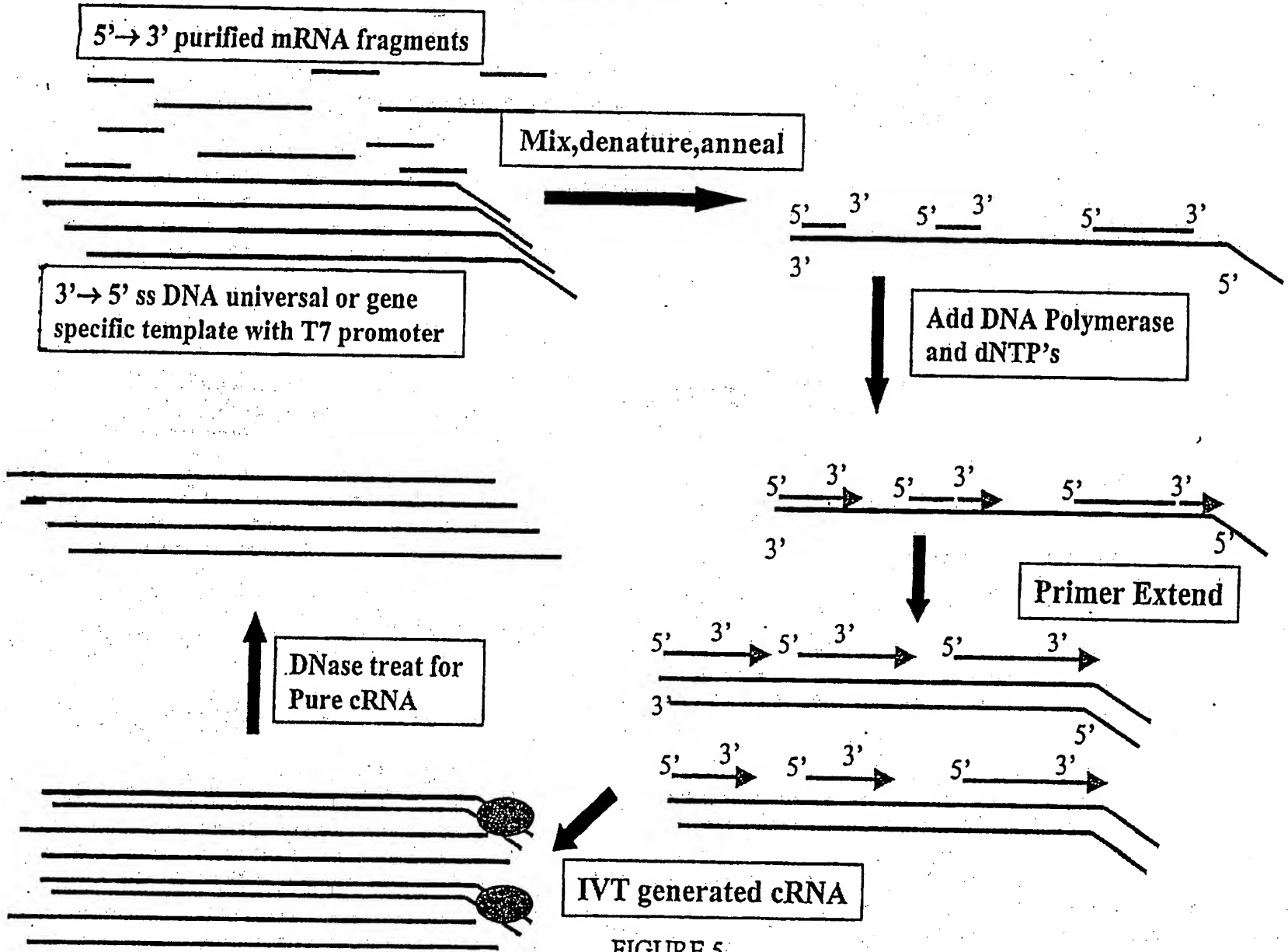


FIGURE 5

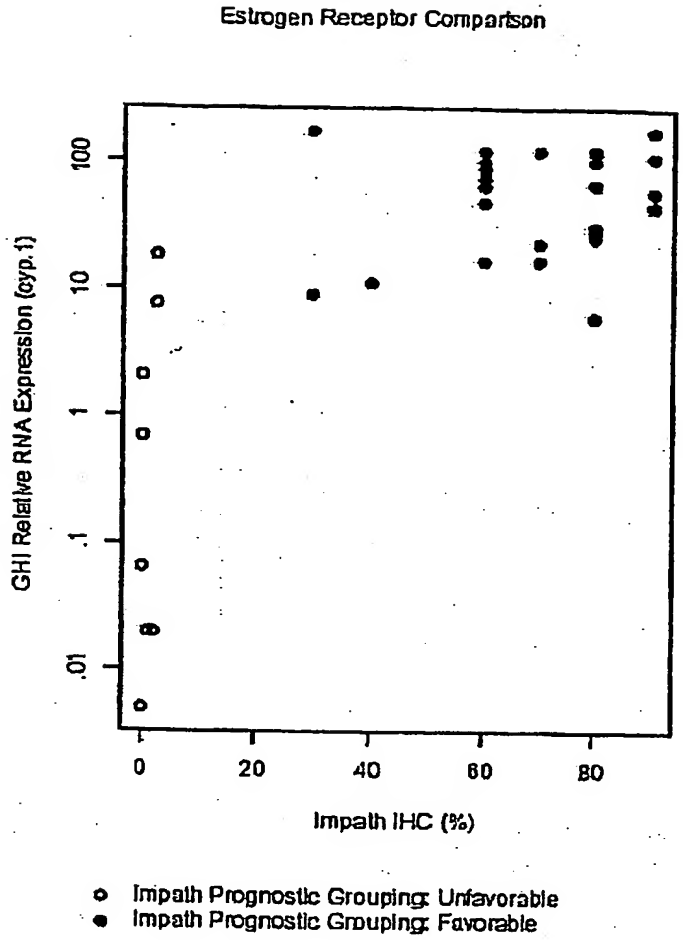

Estrogen Receptor Comparison
GHI Relative RNA Expression (cyp.1)
100
10
1
.1
.01
0
20
40
60
80
Impath IHC (%)
Impath Prognostic Grouping: Unfavorable
Impath Prognostic Grouping: Favorable

FIGURE 6

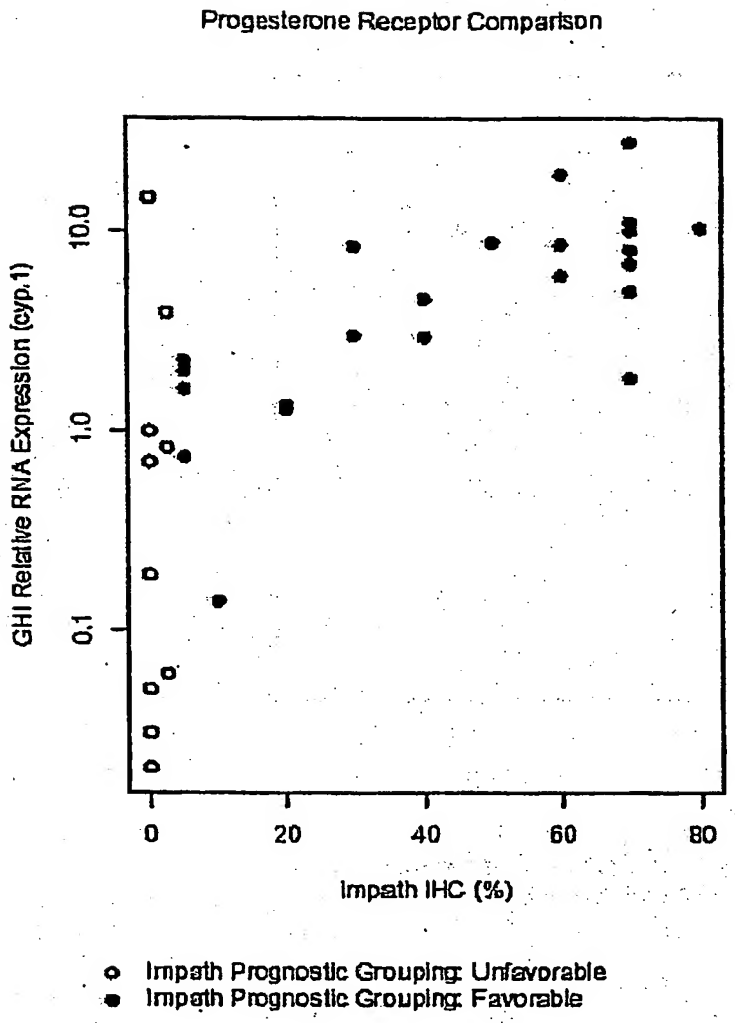

Progesterone Receptor Comparison
GHI Relative RNA Expression (cyp.1)
10.0
1.0
0.1
0
20
40
60
80
Impath IHC (%)
Impath Prognostic Grouping: Unfavorable
Impath Prognostic Grouping: Favorable

FIGURE 7

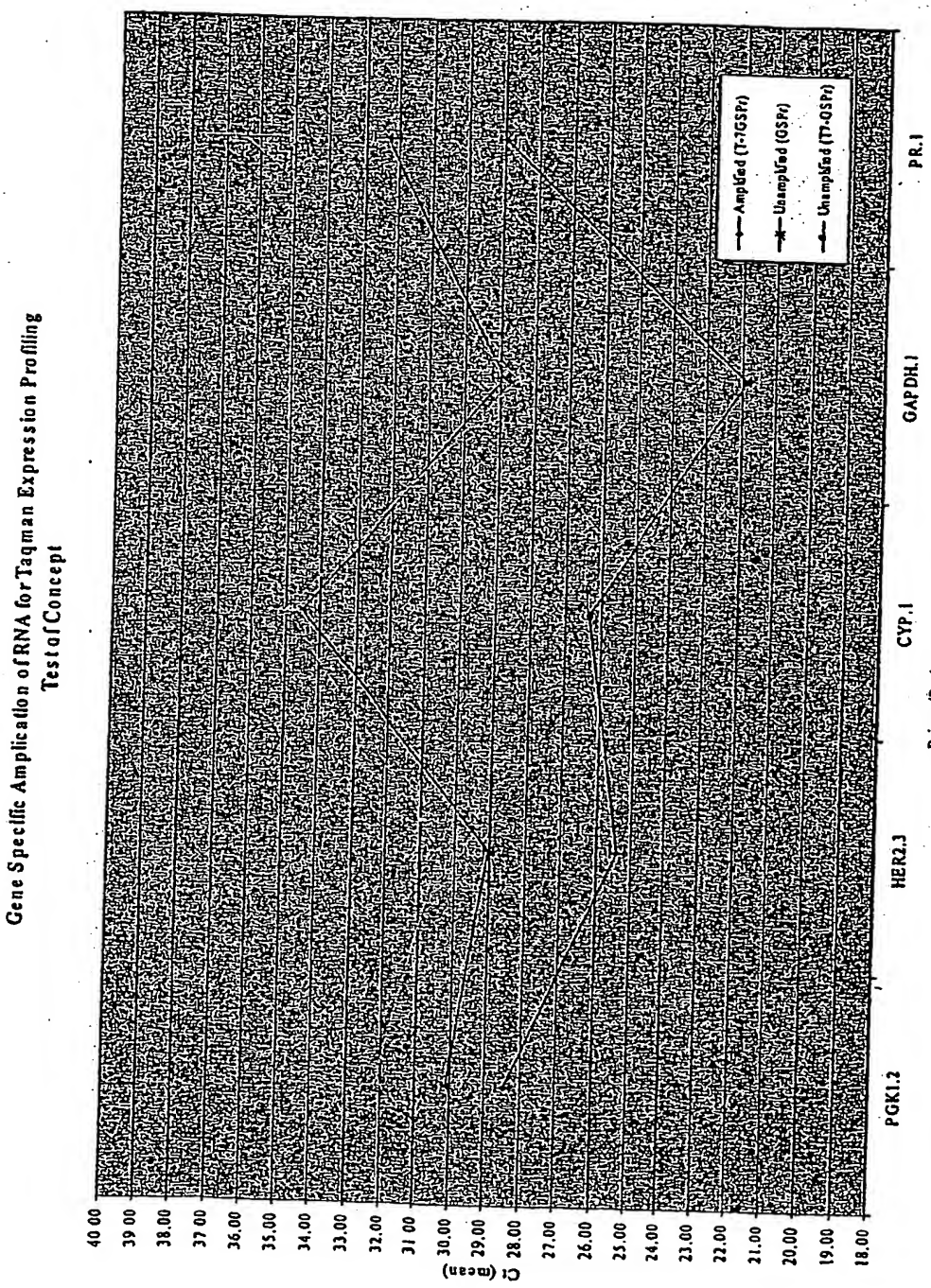
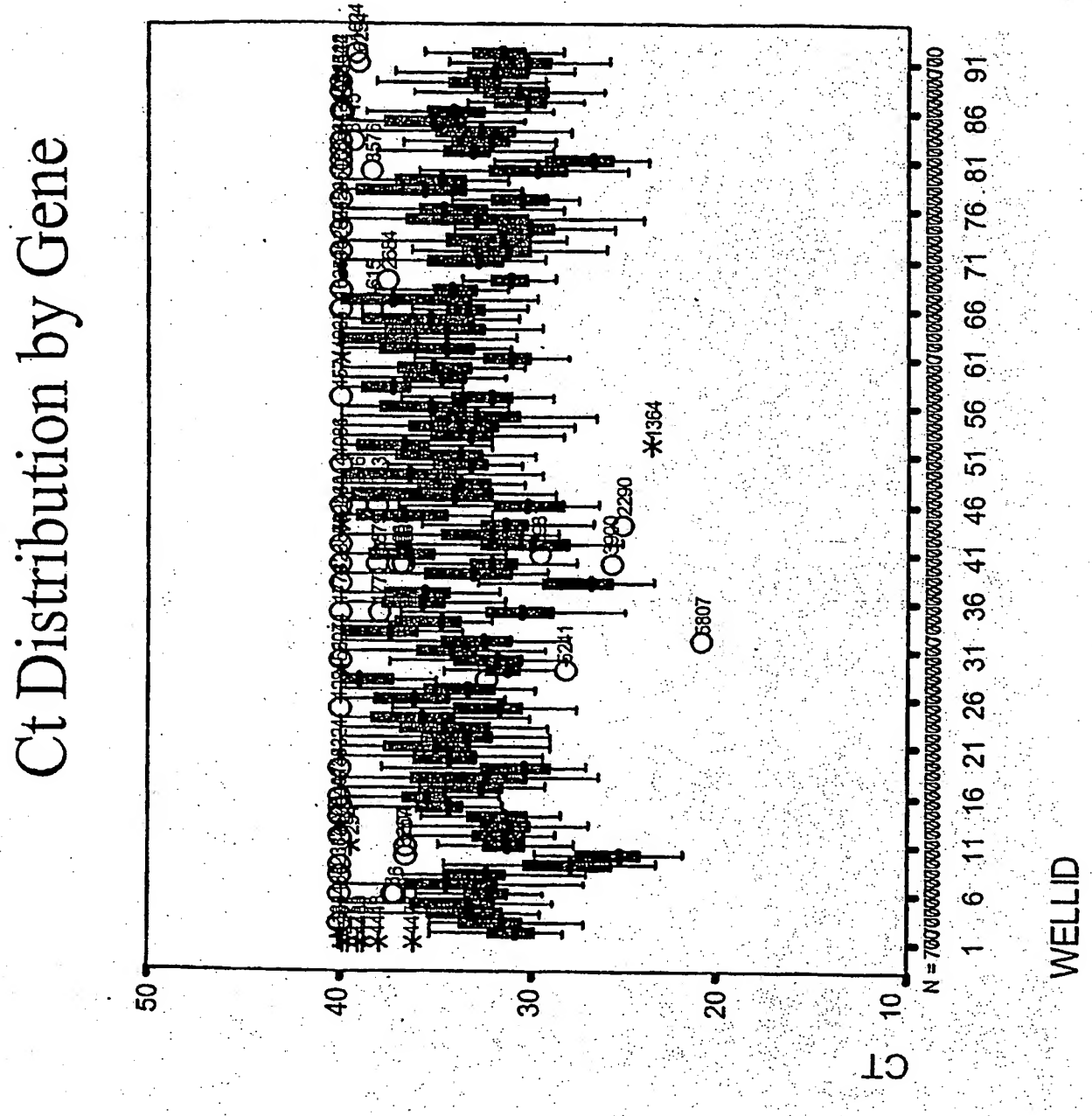

Gene Specific Amplication of RNA for Taqman Expression Profiling
Test of Concept
Amplified (T-7GSPr)
Unamplified (GSPr)
Unamplified (T7-GSPr)
Ct (mean)
40.00
39.00
38.00
37.00
36.00
35.00
34.00
33.00
32.00
31.00
30.00
29.00
28.00
27.00
26.00
25.00
24.00
23.00
22.00
21.00
20.00
19.00
18.00
PGK1.2
HER2.3
CYP.1
GAPDH.1
PR.1
Primer/Probe

Figure 8

FIGURE 9


Ct Distribution by Gene
Ct
50
40
30
20
10
N = 700
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91
WELLID
1364
2290
5807
6241

39740-0001PCT.txt

SEQUENCE LISTING

<110> GENOMIC HEALTH
Baker, Joffre B.
Cronin, Maureen T.
Kiefer, Michael C.
Shak, Steve
Walker, Michael Graham

<120> GENE EXPRESSION PROFILING IN BIOPSIED TUMOR TISSUES

<130> 39740-0001PCT

<140> to be assigned
<141> 2003-03-12

<150> US 60/412,049
<151> 2002-09-18

<150> US 60/364,890
<151> 2002-03-13

<160> 384

<170> FastSEQ for Windows Version 4.0

<210> 1
<211> 18
<212> DNA
<213> Homo sapiens

<400> 1
gtcccaggag cccatcct 18

<210> 2
<211> 19
<212> DNA
<213> Homo sapiens

<400> 2
cccggctgtt gtctccata 19

<210> 3
<211> 68
<212> DNA
<213> Homo sapiens

<400> 3
gtcccaggag cccatcctgt ttgactgcag cattgctgag aacattgcct atggagacaa 60
cagccggg 68

<210> 4
<211> 18
<212> DNA
<213> Homo sapiens

<400> 4
tcatggtgcc cgtcaatg 18

<210> 5
<211> 23
<212> DNA
<213> Homo sapiens

<400> 5
cgattgtctt tgctcttcat gtg 23

<210> 6
<211> 79
<212> DNA
<213> Homo sapiens

<400> 6
tcatggtgcc cgtcaatgct gtgatggcga tgaagaccaa gacgtatcag gtggcccaca 60
tgaagagcaa agacaatcg 79

<210> 7
<211> 20
<212> DNA
<213> Homo sapiens

<400> 7
aggggatgac ttggacacat 20

<210> 8
<211> 20
<212> DNA
<213> Homo sapiens

<400> 8
aaaactgcat ggctttgtca 20

<210> 9
<211> 65
<212> DNA
<213> Homo sapiens

<400> 9
aggggatgac ttggacacat ctgccattcg acatgactgc aattttgaca aagccatgca 60
gtttt 65

<210> 10
<211> 22
<212> DNA
<213> Homo sapiens

<400> 10
tcatcctggc gatctacttc ct 22

<210> 11
<211> 20
<212> DNA
<213> Homo sapiens

<400> 11
ccgttgagtg gaatcagcaa 20

<210> 12
<211> 91
<212> DNA
<213> Homo sapiens

<400> 12
tcatcctggc gatctacttc ctctggcaga acctaggtcc ctctgtcctg gctggagtcg 60
ctttcatggt cttgctgatt ccactcaacg g 91

<210> 13
<211> 20
<212> DNA
<213> Homo sapiens

<400> 13
agcgcctgga atctacaact 20